Aalto University
School of Science
Master's Programme in Human-Computer Interaction and Design

Karolina Drobotowicz

Guidelines for Designing Trustworthy AI Services in the Public Sector.

Master's Thesis
Espoo, July 21, 2020

Supervisors: Professor Marjo Kauppinen, Aalto University
Associate Professor Konrad Tollmar, KTH Royal Institute of Technology
Advisor: Sari Kujala Ph.D.
Nishan Chelvachandran, FRSA

Aalto University
School of Science
Master's Programme in Human-Computer Interaction and Design

ABSTRACT OF MASTER'S THESIS

Author: Karolina Drobotowicz

Title:
Guidelines for Designing Trustworthy AI Services in the Public Sector.

Date: July 21, 2020 **Pages:** 120

Major: Computer Science **Code:** SCI3042

Supervisors: Professor Marjo Kauppinen
Associate Professor Konrad Tollmar

Advisor: Sari Kujala Ph.D.
Nishan Chelvachandran, FRSA

Artificial Intelligence (AI) has been a popular topic in different areas of the current world. Thus, it is natural that its use is considered in the public sector. AI brings many opportunities for public institutions and citizens, like more attractive, accessible and flexible services. However, existing stories also show that the unethical or opaque use of AI can reduce significantly citizens' trust in responsible public institutions. As it is important to maintain such trust, trustworthy AI services are gaining more and more interest. This work aims to answer the question of what needs to be taken into consideration while designing trustworthy public sector AI services. The study was done in Finland. The design process was used as a study method and it consisted of qualitative interviews, design workshop and validation with user testing. Altogether more than 30 Finnish residents participated in the study. Currently, there are more positive than negative voices about the usage of AI in the public sector, however, the number of the latter is significant. The most negative voices were coming from older people of low education and from younger AI specialists. Moreover, strong trust exists in the public sector. Nevertheless, citizens are voicing multiple concerns, such as security or privacy. It is important to keep the public sector services transparent, in order to keep trust in the public sector and build trust in AI. Citizens need to know when AI is used, how and for what purpose, as well as, what data is used and why they receive specific results. Citizens' needs and concerns, as well as ethical requirements, ought to be addressed in the design and development of trustworthy public sector AI services. Those are, for example, mitigating discrimination risks, providing citizens with control over their data and having a person involved in AI processes. Designers and developers of trustworthy public sector AI services should aim to understand citizens and ensure them about their needs and concerns being met, through the transparent service and the positive experience of using the service.

Keywords: artificial intelligence, public sector, trust, trustworthy services, transparency, guidelines, design process, citizens

Language: English

Acknowledgements

First of all, I want to thank my Aalto supervisor, Marjo, who patiently helped me with structuring the thesis and getting the best out of it. I cannot skip in thanks also Sari, who gave me great advice on how to write a better thesis and Konrad who tremendously helped me from the KTH side. Many thanks to Nishan, who paid a great double role of being the industrial advisor, when suggesting interesting reads and connecting with relevant people, and a thesis one, giving advice on what to focus on.

I am also sending thanks to Meeri Haataja for starting the Citizen Trust Through AI Transparency project and inviting me to it. I wouldn't work on this great topic if we didn't meet at the event about data ethics. Thanks also to the whole Saidot and Citizen Trust Through AI Transparency project team who created a great atmosphere and taught me a lot. And of course, thanks a million to all the study participants, who not only helped with their participation but also shared their interest and enthusiasm in the study itself.

Personally, I would like to thanks all my friends with whom we went together through the thesis writing struggles, and to those who patiently understood my lack of time. Big thank you to Lena, for giving me great feedback and becoming my opponent in the last moment. Thanks also to my mum, for reasking me all the time how is my thesis going and for finding a good side of the corona-crisis, that is, that I can finally sit at home and write the thesis. Last but not least, thank you, Dima, for surviving.

Espoo, July 21, 2020

Karolina Drobotowicz

Contents

Chapter 1

Introduction

This chapter introduces the reader to the topic of this master thesis. The first section 1.1 presents the background work and motivation for choosing the topic of this thesis. Next, the research questions that lead this work are presented in section 1.2. The scope of the thesis is described in section 1.3 and the structure of it in section 1.4.

1.1 Background and motivation

Recent advances of Artificial Intelligence brought again more popularity to this topic, after somewhat slow progress during AI winter [1, 2]. Despite this recent focus on AI, there is no one definition yet that majority would agree on [1, 3, 4]. As the context of this study is placed in the Finnish public sector, I will provide here a definition from a Finnish document [5]:

> [...] artificial intelligence refers to devices, software and systems that are able to learn and to make decisions in almost the same manner as people. Artificial intelligence allows machines, devices, software, systems and services to function in a sensible way according to the task and situation at hand.

The number of applications of AI is rapidly growing, therefore gaining also the interest of governments and public organizations [6–8]. There are multiple arguments provided, for why AI would be useful especially in the public sector context. Sun et al. [9] argue, that AI, being more flexible than previous automation technologies, is better suited to the public sector, where environmental settings are constantly changing. Other arguments are that without modern technologies, the public sector is less satisfying than the private one [10] or that AI can lower the administrative burden and take on mode complex tasks, that would enable government workers to focus more on citizens needs and lower the corruption [10, 11]. European Commission sees that AI could be used for services that could serve citizens 24/7 in more agile, accessible and faster way [7].

There are already cases existing of AI usage in the public sector. In the USA machine learning was used to recognize the handwriting on envelopes since the late 1990s [10]. More generally, it has been used in education systems, social policies or health inspections [9]. In Finland, the Aurora AI project is under development, that would become a 24/7 available interface between the citizen and many public services [5, 12]. In fact, Finland seems to be motivated to create "world's best services" [5] with the use of AI. Their general vision from the document from late 2017 document [5] is:

> In another five years time, artificial intelligence will be an active part of every Finn's daily life. Finland will make use of artificial intelligence boldly in all areas of society - from health care to the manufacturing industry - ethically and openly. Finland will be a safe and democratic society that produces the world's best services in the age of artificial intelligence. Finland will be a good place for citizens to live and a rewarding place for companies to develop and grow. Artificial intelligence will reform work as well as create well-being through growth and productivity.

However, currently available services are not always beneficial for society. AI Now report [6] mentions that multiple of deployed automated decision systems are untested or poorly designed, are therefore often redound to misleading results or illegal violations of civil rights. For example, the report mentions cases of cancelling thousands of visas due to system error, unsafe and incorrect cancer AI recommendations or automated decrease of social help allocation without any explanation or possibility to contest it. Moreover, as one survey from 2019 [13] shows, the majority of citizens are not aware of AI being used, neither they are prepared for it. In 2019 AI Now Institute published another report with all cases of automated decision systems used in US public administration [14]. Most of those examples came as an unpleasant surprise to citizens of New York, decreasing the trust of citizens [15].

All those cases show how important applied ethics are in the development of such services. Experts agree that it would help in minimizing negative outcomes [16]. In fact, when users perceive ethical standards of the service as low, the image of the provider may be damaged [17]. Even when it would be service developed by the third party, it is a public organization that would be held responsible by their failings [6]. In summary, creating ethical AI services is needed and it confers a dual advantage for the public sector: from one side it enables to identify and leverage new socially acceptable opportunities, on the other side it helps in preventing costly mistakes [1, 17, 18]. The need for ethical AI is reflected also in the start of new organizations like AI Now Institute, or big grants for this field of study from MIT [16].

Making AI services ethical, helps in building citizens trust to the provider, here public sector [1, 3, 17].Trust can be defined as "as the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" [19]. Trust building to the automation is said to

follow a similar process as to other humans, however, it is not totally alike [19]. Arnold et al. [20] states, that the level of public trust to AI is yet lower than to other technologies, due to less knowledge available and bigger complexity of it [3]. However, some say that it is not AI that should be trusted, but rather the organizations and regulators who are responsible for it [21], specifically the impersonal institution rather than people behind it [22]. Last but not least, trust is said to be dynamic [23], non-binary and context-dependent [24]. For example, a citizen can trust the organization but might not trust the system they use, or vice versa. Talking about trust to institutions or systems, often the adjective "trustworthy" is used. It is used as a measure of how much factors influencing the trust are met [25, 26].

Trust in digital services, therein those with AI is vital for citizens to use them [3, 27] and for society to keep on developing and deploying AI systems [28]. Moreover, trust in the system can also increase the productivity of using it [29]. People are likely to resist technology which they do not trust, even if it promises vast economical or social benefits [30]. Hence, current European politics are focusing on scaling the trustworthy AI [28]. Specifically, they also address the great opportunity of European public sector to play a significant role in uptaking, adopting scaling trustworthy AI [7]. As motivation, they mention that it can lead to new opportunities for research and entrepreneurship, leading to responsible and welfare-enhancing innovations.

Furthermore, trust in the public sector is of the same value, as it is positive for economic, social and psychological well-being [31]. Similarly to the trust in AI, here also lack of it can make citizens resist usage of public sector services or even actively oppose its regulations [22]. On the other side, trust to the public service increases its efficiency and reduces complexity [22]. Nevertheless, it is suggested that some level of distrust is healthy and needed to maintain administrative accountability [32].

To ensure ethical and trustworthy public sector AI services there is a need declared for principles and guidelines [16]. The absence of such can be reflected in the uncertainty of the technology [16]. There are already existing initiatives, directives and guidelines that aim to help in creating trustworthy AI systems [17], public sector e-services or usable digital interfaces [33]. However, as Rostlinger and Croholm hypothesise, design guidelines should be as much context-based as possible, in order not to be perceived as too superficial [33].

Moreover, existing guidelines and research about ethical AI systems are often the results of discussions with industry and academic expert stakeholders, rarely including citizens needs and voices [17, 30]. In the research for explainable machine learning and AI the focus is on technology, rather on usability for end-users [34]. Since the development of sociotechnical theory, we know that efficient systems need to have both technology and users considered during the design and development [35]. Therefore, there is a need stated for understanding public concerns and needs, as well as for including citizens in the public sector AI services development [2, 10]. Last but not least, the final report

of Finland's Artificial Intelligence Programme 2019 [12] states that the already existing trust towards the public sector obliges to them to actively understand the prerequisites for trust and ensure human-centric operations.

This thesis aims to provide guidelines for the public sector that would help them in providing trustworthy AI services, hence building the trust of citizens to AI and the public sector. They are built based on the extensive literature review and the empirical design process with the participation of Finnish residents. The latter process was done as a part of "Citizen Trust Through AI Transparency" [36], organized by the company called Saidot, conducted together with three Finnish authorities: Siitra, Ministry of Justice, Kela; and two Finnish cities representatives: Espoo and Helsinki.

1.2 Research questions

The main research question of this master thesis is as follow: **What needs to be taken in consideration while designing trustworthy public sector AI services?**.

This question is accompanied by four more detailed question, that will lead the focus of this thesis:
RQ1: What are the current attitudes and concerns of citizens towards the use of AI in public sector services?

RQ2: What information about public sector AI services is needed to be transparent for citizens' trust?

RQ3: What factors can affect citizens' trust in AI services of the public sector?

RQ4: What should be included in guidelines for trustworthy public sector AI services?

1.3 Scope of the thesis

The process of answering the research questions is set in the Finnish environment. To be more exact, the empirical work of this study is based purely on interactions with over 30 residents of the Metropolitan Area of Finland. While for quantitative studies 30 might seem a low number, in case of qualitative studies present in this thesis, this is a sufficient number of participants. Regarding the literature review, only a part of it consists of studies done with Finnish or Nordics recipients and Finnish documents, due to the scarcity of those.

By the public sector AI services I mean services provided by public institutions for citizens, that are using AI systems. The examples of such services

could be health-condition predictions, assistance and decision making in applications for social benefits, education or immigration impact assessment on local society or economy. Nevertheless, due to the novelty of the topic, the reviewed literature also includes different AI-related technologies, such as automatic decision systems or machine learning.

The empirical study presented in the thesis was a part of the "Citizen Trust Through AI Transparency" [36], organized by the company called Saidot. The citizen interactions were planned, organized and conducted in collaboration with the Finnish design lead from Kela and graphical designers consultants. With the former, we organized and conducted interviews and user testing. The help of the Finnish designer was especially needed when conducting the interviews with people who were not comfortable speaking English. With the graphical design consultants, we designed and produced the public sector AI service prototype. The study participants were permanent residents of Finland. In the following study, they are sometimes also called as citizens.

This thesis brings three main contributions. The first is the analysis of the current attitudes and concerns of Finnish residents towards public sector AI services. The second is the understanding of what transparency means for citizens and how would they like it to be. The third is the set of guidelines on what to include in the design and development of trustworthy public sector AI services. Those guidelines are based on citizens needs and concerns, as well as, on expert opinions. The practical outcome of the first contribution is in the form of the personas (appendix D), while the practical outcome of the second and third are grouped in guidelines document (appendix F) and visually presented as a service prototype (appendix E).

1.4 Structure of the Thesis

The structure of the thesis is as follow. Firstly, the methods used for the literature review and empirical study are presented in chapter 2. Next two chapters 3 and 4 contain the results of this thesis. Table 1.1 presents which of the literature review and empirical study sections answer on which of the research questions. During the each of empirical study parts, that is interviews, design workshop and user testing, multiple research questions were tackled with different focus. Hence, there is no clear division between the section and research question. Next, the answers to the research questions are discussed in chapter 5. Last, the conclusions are presented in the chapter 6.

	LR 3.1	LR 3.2	LR 3.3	LR 3.4	E: Int 4.1	E: DW 4.2	E: UT 4.3	E 4.4
RQ1	**X**				**X**	x	x	
RQ2		**X**			**X**	**X**	x	
RQ3			**X**		**X**	x	x	
RQ4				x		x	x	**X**

Table 1.1: Which section of the thesis answer on which research question. Legend: **X** - answers on the big part of the question; x - answers partially; LR - Literature Review; E - Empirical study; Int - Interview; DW - Design workshop; UT - User testing

Chapter 2

Methods

This chapter presents the methods used in this study. Firstly, it introduces how the literature review was performed in section 2.1. Next, it presents the process of the empirical study in section 2.2, with its separate parts in according subsections. The processes of literature review and empirical study started at the same time, however, the former was finished after the empirical results were produced.

2.1 Literature review

The literature review started with the beginning of the "Citizen Trust Through AI Transparency" project in 2019 and continued until May 2020. In the beginning, the reviewed materials were the one suggested by people involved in the "Citizen Trust Through AI Transparency" project. They were actively sharing not only scientific papers but also technical reports, online guidelines or directives, such as "Directive on Automated Decision-Making" [37], "AI Now Report 2018" [6] or "Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems" from IEEE [38].

After the empirical study, the materials were searched for using the scientific papers search engine, called *Google Scholar*. I used several different search prompts, that was the combination of following keywords: *Public Sector, AI, Artificial Intelligence, Interface, Design, Guidelines, digital, systems, services, recommendations, design guidelines, ethical, transparency, trustworthy*. The delay with the literature review was caused by time restrictions. On the bright side, multiple relevant new materials were published in the time of delay, such as "Designing Explanation Interfaces for Transparency and Beyond" from 2020 [39] or "Ethical framework for a fair, human-centric data economy" report from October 2019 [35].

Materials were chosen firstly based on their relevance from the title and the abstract. Next, the conclusions were assessed to validate whether the material can indeed be useful. Chosen texts were read and most important quotes and notes from them were saved in the text document. Those quotes and notes were grouped into four different sections relating to the research questions.

When this step was over, the notes inside each group were analyzed and clustered based on the topic they were stating.

2.2 Empirical study

This section presents the methods used in the empirical study. In the first subsection 2.2.1, a reader can see the overview of the design process of guidelines. The following sections presents accordingly how the interviews 2.2.2, design workshop 2.2.3 and user testing 2.2.4 were organized, conducted and analyzed.

2.2.1 Process

The process of the empirical study was inspired by the Double Diamond method launched by the Design Council in 2004, right now being "world-renowned with millions of references to it on the web" [40]. Double Diamond consist of four main parts: Discover, Define, Develop, Deliver [40]. Those are compared below with the process used in this thesis, which is presented in the chart 2.1. The Double Diamond method was, for instance, successfully used in designing environmental sustainability strategies in 2014 [41].

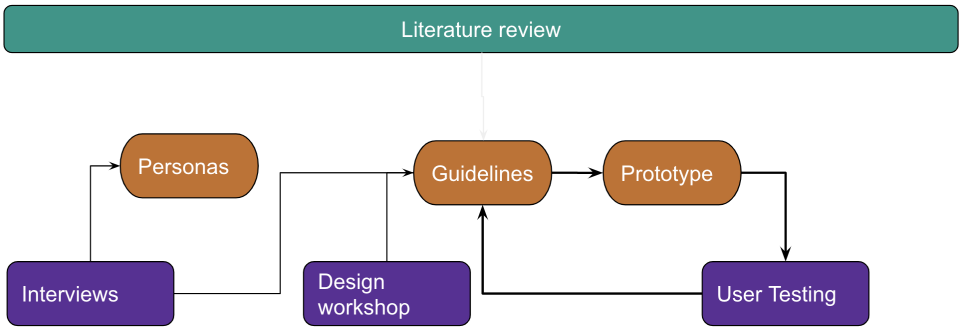


Figure 2.1: The design process of guidelines.

The first iteration in the Double Diamond method is called: Discover. There, it is suggested to research and understand the real problem [40]. Hence, the empirical study started with a series of qualitative interviews with Finnish residents, described in more detailed in subsection 2.2.2. The aim of those interviews was to understand the state of current knowledge, concerns and needs of citizens towards the AI used in the public sector.

The next step in the Double Diamond is Define. During it, designers should aim to define the challenge based on the results of the first step [40]. This step was done through the interview analysis, described in detail in subsection 2.2.2. During that step, the direction on trustworthy AI services was taken. It also resulted in creating personas.

The third step of the Double Diamond is Develop. There it is suggested to provide several different solutions to the defined challenge [40]. It is also suggested to involve a range of different people in designing such solutions, that is to co-design with them [40]. The time scope of the project didn't allow us to provide more than one solution to the challenge, unfortunately. However, the step of co-design ideation with Finnish residents was performed, in the form of a design workshop. This part is described in subsection 2.2.3. The aim of the workshop was to brainstorm together on how trustworthy PS AI services could look like. Based on the knowledge from there and interviews, the first draft of the guidelines was created.

The last step of the Double Diamond is Deliver and it involves testing solutions and developing the best one [40]. Since there was only one solution developed in the third step (guidelines), in this study in the last step we focused on improving it with user testing. However, we realized that the document with guidelines would not be a good material to test with citizens. Therefore, we decided to develop a prototype of the AI public sector service based on guidelines, which citizens could relate to. Hence, Deliver step was an iterative process of following actions. Firstly, we created a public sector AI service prototype based on the created guidelines. Next, we tested the prototype with citizens. Based on the received feedback, we updated the guidelines and accordingly updated the prototype again. The process of updating the prototype and testing it was repeated twice and is described in more detailed in subsection 2.2.4.

2.2.2 Interviews

The empirical study process was started with the series of qualitative interviews. Three goals of them were to:

- Check how much do people know (e.g. How their data is used? What is AI?).
- Understand citizens attitude to AI, especially in regards to AI in the public sector.
- Understand people needs in the cases where AI is used in the public sector (e.g. If they want transparency? How much information?)

. For reaching those goals, we agreed to prepare semi-structured interviews, which provide a structure (e.g. outline of questions or topic to ask) that doesn't need to be strictly followed [42]. Based on the researches [42, 43], we found that semi-structured interviews can provide us a good balance between getting in-depth answers and spending less time on them, comparing to those structured or not structured interviews. The questions used in interviews are attached in the appendix A. The interviews with citizens were prepared, piloted and performed with the design lead from Kela.

We aimed to have 20 interviews with as representative groups for future users of AI public services. Therefore, we decided to interview four different groups of people differentiating on two axes: education level (academic and less) and age (under and above 30). Moreover, plans were to talk with people educated in fields connected to AI and not. All of the participants needed to be either Finnish citizens or live in Finland for 3 years or more. As I cannot speak Finnish, I was conducting the interviews only with those participants who felt comfortable with English (around half). The other part of interviews was led by the designer from Kela in Finnish.

The structure of the interviews was as follows. Firstly, after basic demographic information, interviewees were asked about their general knowledge of data and AI. Therein, we asked for their current attitudes to the private and public sector data usage and how do they understand AI. The second section was focused on the use cases. Every participant was given between 2 and 4 use cases and after each of them asked a few questions regarding their first feelings, eventual concerns and need for clarifications. The third and the last part was to ask a few closing questions focused on using AI in public services. The interview was planned for 45 minutes and participants were offered one movie ticket for participation. Interviews were audio-recorded.

The use cases are attached in the appendix C. They were chosen is such a way to cover different sectors of AI usage like AI assistant, impact assessment, decision making and future predictions. Those sectors were chosen based on discussions with AI experts as well as Public Sector representatives. Moreover, we aimed to make as easy as possible for interviewees to relate to the presented situation. Therefore, for one of the use cases, where we talk about the prediction of future social exclusion, we use two examples: either with grandmother or grandchild. Those were given out to participants depending on their age, so younger participants were given an example with the grandmother as they might not have yet associations with having children in school. Furthermore, those cases are on purpose very scarce in any information. They are often lacking reasoning, purpose or information about the data source. The reason for that was to nudge participants to say what is being missed the most. Mentioned use cases were given out to participants in the counterbalanced order, to avoid the bias.

The interviews were analysed with the grounded theory approach, due to the novelty of the researched topic of AI use in the public sector, as suggested in a few studies [42, 44]. No hypotheses neither codes were stated prior. As a first step, interviews recordings were transcribed and uploaded to the Atlas.ti tool. Later, while reading the transcriptions, important citations were selected and coded. The codes were being generated and updated during the whole process. At the end, each code was consisting from two to four parts representing its location, type and meaning, such as: UC1_feeling_positive_excited. Later on, codes were checked with the demographics of participants to understand existing dependencies. They were also clustered by topics to understand the repetitive patterns and summarize citizens attitudes, concerns and needs.

2.2.3 Design workshop

Design workshop was a second step of the empirical study, after the interviews and its analysis. Its goal was to engage citizens in creating the interfaces of trustworthy AI services by a co-design ideation session. The method for the workshop was inspired by the ideation methods described in the book of Michanek and Breiler [45].

For the workshop I invited eight Finnish citizens and residents. Six of them also took part in the interview. The participants knew only the topic of the workshop before coming, which was about trustworthy AI services of the public sector. The workshop was planned to last two hours and participants were offered two movie tickets for the participation. Some beverages and snacks were served during the workshop.

The workshop started with a warming up and ice-breaking game. Participants were asked to line up in the given space from the lowest to the highest level of how their day went, how much do they know about AI and how much would they trust AI. Later, participants were put in groups of 2, 3 and 3 and sat down in indicated places. Each group received one type of AI usage in the public sector. In the groups, people were asked to acknowledge given materials, discuss it and then save the results of the discussion in the writing, notes or drawings. Each group was given blank A3 papers, post-its, pens and colourful markers. The main area of the questions asked to each of the cases was how to make the following case trustworthy.

After the first round of the ideation, participants were asked to cover their results and change places. Both the case they worked on and groups were changed. The motivation for performing those rotations was to increase the opportunities for innovative and exploratory approach while discussing with people of different perspectives, as well as to minimize bias and stagnation. In the second and the third round, participants were firstly doing the same actions as in the first one. After around 10 minutes though, they were asked to uncover previous group(s) results. Then, they could either get inspired by previous ideas or comment on those.

Used cases can be seen in appendix C. Each of them consists of short introduction, information about possible input, process and output, example case and questions. The first case is related to the decision making AI, for example where a person looking for a student flat would be assigned one automatically by AI. The second was about the impact assessment by AI, for example where government would be willing to measure the impact of the education in Finland on Finns' well-being by tracking their data, like health or income. The third use case was about predictions done by AI, where artificial intelligence could be used to provide the prediction about possible disease risk for you, based on your work and family health data.

The workshop was ended with the whole group discussion. Firstly, each of the AI usages in public service was discussed. Specifically, participants were asked for their most important insights from the brainstorming stage. Later,

we discussed the whole topic of AI use in the public sector and the workshop itself.

Similarly to interviews, design workshop analysis was also done in the grounded theory fashion. However, no digital tools were used for analysis there, as the amount of collected materials from the interviews was smaller than from the interviews. The main part of the analysis was affinity mapping with use of post-its, where the post-its were grouped based on topic resemblance.

In the first step of the workshop results analysis, each use case was analysed separately. The topics that appeared there were relating to the questions are post-its are answering on, like what should be included in the interface or when and how should a person be informed. The clusters created based on those topics were saved in the text document. As the next step, all post-its from all use cases were mixed together and clustered based on the topic of the context they consist, such as transparency, data sharing or human involvement etc. Those clusters were also saved to the text document. At the end, there were also comments from the discussion added to each of the sections. In summary, the results presented the information on how transparency should be performed in different public sector AI services, as well as, what needs to citizens have towards such services.

2.2.4 User testing

The user testing was the last step of the empirical study. It aimed to understand whether the information and ideas gathered in interviews and design workshop are correctly understood by us and complete. For reaching this goal, I first created the first draft of the guidelines for the design of trustworthy public sector AI services, that addressed all needs, transparency vision and concerns that were listed in the previous interactions with participants. However, together with other project stakeholders, we realized that testing guidelines with Finnish citizens might be not successful for our needs. It might have been difficult for people, who are not designers, to understand, relate and check such guidelines. Hence, the idea came for creating the AI public sector service based on the guidelines.

We decided to create a digital prototype, that would be easily relatable to a diverse group of Finish residents. In the fake service, coming from the public sector, citizens would be offered predictions of their possible health issues in the future. The visual, web-based prototype was created by the external partner of the project and is presented in appendix E. It was designed following the first draft of the guidelines. In the beginning, the prototype was divided into three stages: application, where clients could choose which data they want to share; waiting, when the data was being processed; and results with the possibility of sharing the results with other organizations. Later, the informative stage was added.

The user testing was done in three iterations. The first one was piloting

and is not used in the below results. That stage, however, helped us to find some issues in the prototype, which upon fixing made the prototype more relevant for citizens. The next two iterations started with the user testing of the prototype with three to five participants per round. Next, the results of the testing were analysed. The final step of the iterations was to update the guidelines and prototype, based on the analyzed feedback from the testing. All iterations led to the state, were a prototype, and therefore guidelines would be approved by citizens.

Chapter 3

Literature Review

This chapter presents the results of the literature review. Its sections are representing answers to four helper research questions. Section 3.1 tells about current attitudes to and concerns about the public sector, artificial intelligence and AI used in the public services. Next, the needed transparency, that is service information that needs to be visible for citizens, is described in section 3.2. The factors needed for the development and operations of a trustworthy public sector, AI and PS AI services are described in section 3.3. Finally, the last section 3.4 presents the current state of reviewed knowledge about what is needed in guidelines.

3.1 Attitudes to and concerns about the use of AI in the public sector

This section presents the results of the literature review on the attitudes and concerns about the use of AI in the public sector. It starts with the analysis of attitudes separately to the public sector and AI. Due to the local dependency of the citizens trust to public organizations, the first subsection 3.1.1 relates only to the state of trust of Finnish citizens to its public sector. In the next two subsections 3.1.2 and 3.1.3, the current attitudes and concerns towards solely AI are presented. Finally, the attitude and concerns together to the AI used in the public sector are grouped in subsection 3.1.4. All concerns are grouped in the table 3.1.3.

3.1.1 Trust in Finnish public sector

When living in Finland one can see, that trust is an important factor for Finnish citizens. In fact, the confidence in public organizations is what makes Finland strong [32] and no other EU country ranks higher in the level of citizen trust than Finland [12]. In 2008, Salminen and Ikola-Norrbacka [32] run a citizen survey with almost 2000 respondents where they measured trust levels in different public and private organizations. As their analysis shows, especially

trust in the public institutions is ranked high - on average 80 % of respondents agreed that they have trust to the public sector and societal organizations. When looked into details of the survey, especially police, education system and military were highly ranked. The government and politico-administrative institutions were ranked notably lower, however still positively [32].

Salminen and Ikola-Norrbacka [32] also researched on the current opinion of different features of public administrations. They report about only two features that were ranked positively: suitable behaviour of public servants and accessible application forms. Close to the neutral point were statements about the clarity of the language and processes. Ranked as the worst was a fact of delays in the processes.

3.1.2 Current attitudes to AI

As to my knowledge, the most extensive study on the public attitude to the AI was done by Fast and Horvitz in 2017 [2]. They looked for long term trends in public perception of AI-based on articles published in the New York Times between January 1986 and June 2016. One of the core findings was that there were a two to three times more of optimistic articles than pessimistic, no matter how much publicity AI had in general. However, it was also accented, that concerns like the existential fear or worry about the jobs are as well growing in popularity in the last years. Another study from the UK, however, mentions, that citizens are yet not aware of AI being used [30]. Only one third would say that AI is used in different decision making and around one-tenth that it is being used in workplace and justice systems.

Three other studies focused on testing the adoption and perception of automated systems in real-life situations[46–48]. The first [46] tested how people are adopting algorithms in tasks with possible collaboration between human and AI. Their results indicate that people are more demanding to automated systems than to humans. While we are able to forgive people for occasional mistakes, even faltering of algorithms can make us less likely to use it, which keeps being true, even when the system actually outperforms humans [46]. That was confirmed also in a study conducted by Dietvorst et al.[47]. They found out that people believe that algorithms cannot learn from their mistakes and therefore are easily becoming much less trustworthy after making errors.

The third study focused on the perception of management decision made by AI in comparison to those made by the specialist [48]. The main outcome of the study was that the trust in algorithms is task-dependent. In mechanical tasks, like work assignment or scheduling, participants found decisions made by algorithms and human specialist equally fair and trustworthy. Those made by algorithms were described as efficient and objective. However, in human tasks, such as hiring or evaluation, decisions made by algorithms brought more negative emotions, felt less fair and gained less trust. As a reason, participants of the study mentioned algorithms' perceived lack of intuition and subjective judgement capabilities. Moreover, the experience of being evaluated

by machine felt dehumanizing, in contrary to the feeling of being appreciated when it is a human specialist performing the evaluation. Only a small group of participants mentioned algorithms as fairer in making this type of decisions, due to their lack of human bias or favouritism. That was partially contradicted with the citizens' jury conducted by RSA Forum of Ethical AI [30]. There, participants also agreed that they are open for the use of mechanical tasks, but as examples gave evaluative tasks, such as deciding about the raise or promotion. As a reason, they started to be attracted by the unbiased assessment of their performance.

The favourability of algorithms also depends on the context of the culture one is a part of. Nitto et al. [49] reviewed a survey, which checked attitude to various types of robots amongst residents of three different countries: the USA, Japan and Germany. For example, the favorability for testing self-driving cars on US roads is in general positive. In more detail, Japan has the least people who do not like it (11 %), the USA has the most people who are extremely favourable (26 %) and Germany has the most of those opposing (27 %). The other system, AI phone operator, on the other hand, was scored mostly positively, especially in Japan.

Few studies yet mentioned which groups tend to trust algorithms more. According to Lee and See [19] higher complacency makes people trust automation in a smart way, that means, the trust is more conditional, aware and sensitive of possible failures that might happen. In the study of Alexander and Blinder [46] it was discovered that the more educated participants were, the better algorithm adoption was happening. They also found out that women were trusting in algorithms more, but it might have been an artefact of women having generally better education. In the other literature review, it was summarized, that trust in automation is affected by human traits such as age, gender, ethnicity or personality, however, more research is needed to state concrete results [23].

3.1.3 Current concerns about AI

There are various concerns appearing in the researches about AI implementation, all of them are grouped in the table 3.1.3. Fast et al. list different worries that were the topic of articles of New York times in the last 30 years [2]. The most frequent and growing ones are worries of humans' loss of control of AI, absence of appropriate ethics for AI, and the negative impact of AI on work. Another, frequently mentioned concern is lack of progress of AI (advancing much more slowly than expected), however, it is descending in the popularity in recent years.

Other studies mentioned concerns like: being tracked, not being able to evaluate qualitative features, inability to accommodate exceptions (or treating everyone homogeneously), potential errors, loss of human contact and empathy, misuse, social injustice, bias or threat [2, 3, 30, 46, 48]. The last of the concerns is specifically explained by Elkins et al. [50]. Feeling of threat seems

Concern	Description	AI	AI in PS	No
Loss of control over AI, threat	Situation when AI is behaving unpredictably and uncontrollably, the decisions are not transparent, or when AI would have control over human and its actions are not questioned by experts.	[2, 3, 50],	[4, 11, 13]	6
Bias and discrimination	Social inequality and unfairness caused by AI applications.	[30, 48]	[4, 13]	4
Potential error	Error occurring in the AI system, potentially leading to disastrous outcomes.	[30, 46, 48]	[4]	4
Poor evaluation	AI not being able to evaluate qualitative features, and therefore unable to accommodate exceptions and treating everyone homogeneously.	[30, 48]	[4]	3
Loss of human contact	AI actions being perceived as demeaning and disrespectful, missing human contact.	[30, 48]	[9]	3
Lack of ethics	Absence of appropriate ethical standards for AI, eg. using it for scoring	[2]	[4, 13]	3
Misuse of data	Unethical and non-transparent data sharing between organizations, eg. sharing patients data with commercial insurance companies.		[9, 30]	2
Misuse of AI	Misuse of AI services for wrong actions, like manipulating populations.		[17, 30]	2
Non-moral AI decisions	Whether it is moral for AI to make decisions for human.	[30]	[4]	2
Negative impact on job	Being replaced by AI in a job, hence loosing hob.	[2]	[4]	2
Lack of AI progress	AI science not developing any further.	[2]		1
Being tracked	Feeling of being put under surveillance.	[48]		1
Accuracy	Results generated by AI being not accurate enough.		[13]	1
Capability of PS to use AI	Public sector not having enough in-house knowledge to run AI services.		[13]	1
Underuse of AI	Underusing AI technologies below their full potential, which can lead eg. to significant opportunity costs.		[17]	1
Blaming the technology	The responsibility for any actions and decisions is put only on the system, which leads to accountability dysfunctions		[11]	1

Table 3.1: List of concerns across studies about AI and AI in Public sector, and the number of occurrences.

to appear when expert users receive counter-attitudinal advice. Followingly, that can generate a negative attitude towards the system. Moreover, in the RSA jury [30], citizens also doubted whether making any decisions based only on statistics can be morally acceptable.

Furthermore, Fast et al. [2] analyzed the hopes appearing in New York Times articles. The most often repeated hopes were: a positive impact on work (mechanical tasks are done by robots), decision making (help in making better decisions with AI or expert systems) and entertainment (better games experience, recommender systems). Moreover, following hopes were mentioned twice less but still on the significant level: improvement of education, transportation, healthcare, merging of human and AI. There was no growing tendency discovered in the frequency of articles containing hopes, but it stays on a higher level than the frequency of the ones with concerns.

Two survey studies were found that checked Finnish citizens' attitude to Artificial Intelligence topics [21, 27]. The first focused on how citizens are using different accessible services that use their personal data. The survey was conducted in four different countries: Finland, Germany, Netherlands and France, in each asking around 2000 inhabitants aged 18-65, below summary represents only Finnish results. One of the questions asked was about terms and conditions. There, 37 % admitted to reading those and 39 % said to understand them fairly well. They also researched whether people change settings for two different reasons: personal needs and due to news about leaks. For the former, 33 % said to adjust settings and for latter 27 %. The reasons for that can be feeling it is not important (30 %) or not knowing how to change those (20 %). In general, Hyry [27] mentions that Finland has the lowest percentage of lowering the use of services due to leaks.

Demographic wise, Hyry [27] found out that students read terms and conditions the least often, in contrast to people of vocational or compulsory education level. Privacy settings changes were done most often amongst young adults (18-24 y.o.) and were dropping with the age. It was also noticed that the lack of trust to AI is greatest with senior and senior staff, lower-salaried employees, entrepreneurs and respondents aged 25-34.

In Trust & AI report [21], 412 Finnish citizens were asked about topics like emotions triggered by AI and trust to different usages of it. For the former topic, there were three emotions on a lead in responses: optimism (57 %), doubt (57 %) and excitement (52 %). Fear was mentioned by 18 % and joy by 12 %. When asked about trust to AI in general on a scale from 0 to 10, the average response was 6.5. In the survey, participants were also asked about trust to AI used for making decisions. There around half of the respondents would not trust AI used in the job application process, whether from the perspective of the employer or applicant.

3.1.4 Current attitudes to and concerns about AI used in Public Sector

In 2019 an international consulting company conducted a survey with more than 14 000 internet users around a world to check their attitude to decisions made by AI, especially in public services [13]. The results were optimistic. For 11 out of 13 cases they presented, participants were cautious positive. That is, they were positive about using AI in those cases, however, they voiced various concerns. Two cases that were not accepted by the public were decision making as parole or judge.

On the other hand, research and healthcare seem to be the most supported sectors for AI use [30, 51]. Especially in Finland, people are also content with sharing their data for scientific research purposes [27]. In another study conducted in Nordics, 69 % of participants would trust medical decisions made by AI, 43 % of which only when human would be involved in the decision process [51]. Furthermore, in the study conducted only in Finland, around 66% of respondents would not trust medical advice if their doctors will not recommend it or not tell how the decision was done [21].

The international survey also checked for a correlation between the demographics of participants and their attitude [13]. It found that younger people are the more trust they have towards use of AI. Also, urban residents are on general more supportive towards AI. Moreover, it was discovered, that the support for AI use is bigger in countries of less developed economies, where the level of corruption is higher. On the other hand, Carrasco et al. found out that support for AI used in the public sector moderately correlates with the trust in the government in countries of more stable economies.

In the survey, participants were also asked for their concerns related to AI used in the public sector [13]. It listed worries such as ethical issues being not yet resolved, lack of transparency in decision making, incapability of the public sector to use AI, the potential for bias and discrimination and accuracy of the results and analysis. Another concern was emphasised in the UK's citizens' jury [30]. There, participants were the most worried about data collected by the public sector being used by other organizations, like for insurance claims.

Four other studies presented expert opinions on possible wrong scenarios when AI would be used in public sector [4, 9, 11, 17]. Firstly, they voice the worry of possible misuse of an AI tool and data it uses. Wirtz et al. [4] focuses on possible misjudgements and AI discrimination bias, while Floridi et al. [17] add the possibility of underusing AI tools, which can cause opportunity costs. On the other hand, Sun and Medaglia [9] mentions that citizens who are not convinced about AI use might lose their trust in public organizations. Smith et al. [11] presents two other possible dysfunctions. First, focus on the outputs, where automated systems gain too big authority and therefore its results are not questioned by specialists and the room for flexibility is limited. The second, called blame the technology, drives from the first. There, the responsibility for any actions and decisions is put only on the system, which leads to

accountability dysfunctions.

We can learn about the Finnish-specific attitude to AI used in the public sector from AI & trust report [21]. From one side, 46 % of study participants voiced that they are concerned to use AI for decision making in the public sector (in comparison to 36 % for the private sector). They were most concerned about using it in surveillance and security and the least in scientific researches. On the other hand, when asked about an opinion about applying AI in the public sector, 40 % of respondents actually stayed neutral, 29 % positive and 17 % negative. Asked for specific data that citizens are ready to share with the public sector, the biggest group (37 %) would be good with sharing health data, while the smallest (2 %) with sharing offline relationships.

3.2 Transparency of AI services for building citizens' trust

This section is focusing on the information about the PS AI services that are needed to be visible for citizens to trust it. In other words, it focuses on the level of needed transparency. The term *transparency*, as well as motivation for having it, is explained in the subsection 3.2.1. The next subsection 3.2.2 focuses on what information is needed from the public sector services in general, whether or not they are using AI. From the other side, subsection 3.2.3 is introducing the transparency needed for AI systems in general. Finally, subsection 3.2.4 presents transparency requested for the public services that would use AI. Grouped information needed for trustworthy of the public sector, AI or PS AI services are presented in the table 3.2.2.

3.2.1 Definition of and motivation for transparency

The transparency has two meanings in the AI sector: first, the technical interpretability of AI actions; and second, the justifiability of the AI processes and outcomes, focused on the whole service that would use AI [52, 53]. In this study, I am interested in the second meaning, that is more relevant from the perspective of citizens, as it includes factors such as availability of the information, conditions of its accessibility and relevance for the citizens [53].

There are multiples motives in the literature for having AI systems and services transparent. Firstly, it allows citizens to better understand and AI actions, which followingly leads to their more positive attitude to the system [3, 31]. It also enables citizens to better evaluate AI actions, that helps in preventing risks or misuses [16, 38, 54], encourage to the use of AI [17, 34] or help in remedying bias [55]. Moreover, Turilli and Floridi [53] argue that transparency is a "proethical condition for enabling or impairing other ethical practices or principles".

Most importantly, transparency is needed for citizens to trust and accept the services, as indirectly presented in the above points and directly in several

other materials [3, 20, 23, 34]. Specifically, the public sector is requested to keep great transparency [4, 6, 10, 55]. That is confirmed by studies with citizens done in Finnish [21] and Nordics [32] environments. However, the information presented has to be understandable by its users to avoid disinformation [52]. The research gap for accessible transparency is mentioned by Abdul et al. [34].

3.2.2 Transparency of public sector services

There are several aspects that are recommended for the public sector in the general context in order to keep or grow citizens trust. Rostlinger and Croholm focus on making sure that public sector services are perceivable, comprehensible and well-preparing for any next action [33]. They suggest to include clear information about the purpose of the service (What is the offer? Who benefits from it?) and the overview of the service process. Moreover, in any e-services, the navigation should be clear, before taking any actions users should know what will happen next and they should be receiving feedback from the system about was they and system did. Additionally, Lee adds the need for information about procedures used for any decision making [48].

3.2.3 Transparency of AI systems

Most of the literature about AI transparency are guidelines which are based on expert knowledge. However, three studies were found, where potential users were asked for their needs. Tsai and Brusilovsky performed a design study, where they asked users for their needs in decision-making AI system, with the focus on explanations [39]. First of all, they discovered that for users to call the system transparent, they needed to see a personalized explanation, data sources and process in an understandable manner. When asked for factors, that bring the trust, the answers were focused on clear information about the benefits that service brings, process, explanation and factors used for the decision. However, when asked to prioritize different system factors, users brought slightly different results. As the most important, they chose beneficial functionality, easiness to use and personalized information. The least important factors turned out to be data sources, raw data and the detailed algorithm used. In Finnish Trust & AI report [21] authors found that two factors increasing trust towards AI are: being informed why and how AI is used (purpose and process) and that the use of AI has been ethically certified. In the report of RSA, it was mentioned that for transparency users need to know whenever an automatic decision system was used and what criteria it took [30].

Amongst other studies and guidelines, an explanation was mentioned the most often as the important factor of AI systems [3, 28, 34, 38, 56–59]. An explanation can be defined as making clear why the AI system behaved like it did (therein why it proposed specific decision) [56], moreover, they should include the reasons and criteria for specific outcomes. In the case of black-box (where it is impossible to state exact reasons), it is suggested to communicate

Name	Description	PS	AI	PSAI	no
explanation	Why AI system behaved like it did, therein why it proposed specific decision.		[3, 28, 34, 38, 39, 56–59]	[4, 31, 37, 52, 60]	14
process	How data was collected, used and processed and whether there are people involved.	[33]	[17, 19, 21, 34, 38, 39, 53, 56, 58]	[37, 60]	12
purpose	Why this system exists, what is the offer and who benefits from it.	[33]	[19, 21, 28, 38, 39, 55, 56, 58]		9
used data	What data is used in AI system.		[28, 34, 39, 58]	[4, 10, 61]	7
code	What code or algorithm is used in AI service?		[28, 39, 62]	[6, 11, 37, 62–65]	7
performance	What are the system limitations and reliability, therein output confidence and accuracy.		[19, 28, 38, 57, 66]	[52]	6
data source	Where are data coming from?		[28, 34, 39, 58]	[10, 61]	5
accountability	Who is accountable for service outputs?		[17, 28, 66]	[4]	4
AI presence	Whether AI is used and for what?		[28, 30, 53, 62]		4
procedures	What procedures, criteria or laws were used in AI service actions, therein decisions.	[48]	[30, 39]		3
redress	How can one post a redress towards AI service outcomes?		[17, 28, 66]		3
impact	What is the impact of the AI service on users and society in general?		[28, 66]		2
data sharing	Whether and with whom is data shared?			[4, 61]	2
certification	Whether the AI system is certified?		[21, 27]		2
service stakeholders	What organizations stand behind the system?		[19]		1

Table 3.2: Information needed for ensuring transparency of Public Sector, AI and AI service used in the Public Sector , with the number of mentions counted

other explicability measures, such as system capabilities or traceability [28]. They should be timely and adapted to the user, they "should take the same form as the justification we would demand of a human" [62, 66]. For the motivation, accessible explanation enhance users' satisfaction and increase their trust in the system actions or adherence to its decisions [3, 34, 59]. Explanations are specifically important in systems that may affect human well-being, therein judicial systems [28, 38, 57].

Next three most-often mentioned transparency factors are, how Lee called them, 3p: process, purpose and performance [19]. Process describes how the system operates, therein what actions it takes based on given inputs and what it can do [17, 19, 34, 38, 53, 56, 58]. Moreover, it is advised to include information about algorithms used and give access to the code for relevant authorities for verification [28, 62]. The transparent and understandable process can bring trust and help in achieving users goals [19]. Purpose explains why the system was developed, for what usage and who benefits from it [19, 28, 38, 56, 58]. Transparent purpose also helps in remedying bias in AI systems [55]. Last from the three, performance states about system limitations, reliability, predictability and ability, therein output confidence and accuracy [19, 28, 38, 57, 66]. Moreover, two guidelines recommend to include the information about potential or perceived risks, that is about likelihood and impact of system errors and fairness [28, 66].

Other factors of transparency are: statement when AI is used [28, 55, 62]; accountability [17, 28, 66]; used data and its sources [28, 34, 58]; system feedback [23] and navigation [56]. The first focuses on the principle, that AI systems should never represent themselves as human, but they should be identifiable as such. As part of the accountability, it should be clear who is responsible for specific AI actions and whom to contact in case of the need to redress the output. The third factor states that it should be clear to the user what data is used and how was it accessed. The last two factors focus more on the transparent usage of the system. At all times user should know what is happening in the system and should be aware of the consequences of different actions.

Several guidelines are sharing their opinions on how to form transparency. Most importantly, the language used in the service should be chosen for the end-users, so it is easy to read, interpret, understand and act upon. That concerns service parts such as terms, decisions or information about the system [28, 38, 62]. Moreover, information shared with users should be relevant for the user context and concrete [19, 56]. It should be provided ahead of time [23] and cannot be false, misleading, partial or with inappropriate details [53].

3.2.4 Transparency of public sector services using AI systems

The biggest difference in transparency guidelines for the private and public sector is that it is asked in the latter services to open source the code or algorithms used. In UK service standards, the reason is stated as such: "Public services are built with public money. So unless there's a good reason not to,

the code they are based should be made available for people to reuse" [63]. As other motivation it is stated, that open-sourced code improves public sector accountability, through opening it for public scrutiny [6, 11, 64] and enables the feedback [65]. However, it is noted that the code should not be open-sourced when it presents a high risk if misused [37, 62].

Most of the other factors are mirrored from the private sector. Services should inform about accountable personas [4], the process [52, 60], explanation [4, 37, 52, 60], performance [52] and data [10, 61]. For the explanation in Public Sector context, Grimsley and Meehan add that it is important to mention why others are successful when the user is not [31]. Leslie adds that explanations should be socially meaningful and demonstrating that the system is "ethically permissible, non-discriminatory/fair, and worthy of public trust" [52]. Regarding data, it is important to inform what data is collected and where it goes [4]. UK government adds information such as what personal information is used for; types of data; what, to whom and why is shared; how long data is used; legal background of having specific data [61].

The instructions on how to present system information are also similar to the ones presented in the previous subsection. Firstly, it should be specific, as otherwise it might be not noticed or negatively influence citizens' perception. It also should be easily visible, clear and memorable to increase trustworthiness [22, 60]. Leslie focuses on easy language and advises adding user-friendly explanations of any more complex terms [52]. Next, the public sector should be proactive in initiating communication and sharing information about automated processes [31, 37]. Smith et al. add that effectiveness of transparency depend on factors like completeness and timeliness of the information [11].

3.3 Factors that affect the citizens' trust in AI services

The following section is presenting the factors that are needed in design, development or operations of public sector AI services. The first subsections 3.3.1 focuses on the factors that are requested from any public services. Next, factors needed for trustworthy AI services are presented in subsection 3.3.1. Finally, the factors requested specifically for the public sector AI services are presented in subsection 3.3.3. All grouped factors are shown in the tables 3.3 and 3.3.

3.3.1 Trust building factors to Public Sector

Heintznman and Marson [67] in 2005 did a survey, where they researched on factors influencing trust in the Public Sector. The results were as followed: keeping of promises, word of mouth about services, staff interactions with the clients, learning from mistakes, interest in citizens' view and quality of leadership in organizations. Salminen and Ikola-Norrbacka [32], on the other hand,

what	desc	PS	AI	PSAI	no
Human involvement	Human is involved in the process of auditing, testing or monitoring of AI system. Human decides on their and AI role. Human intervention is enabled whenever needed.	[30]	[3, 17, 38, 57] [28, 62, 66]	[4, 6, 37, 60] [10, 54]	15
Security and reliability	Proactive prevention of system failures. Risk management, ensuring safety and reliability of systems, security of data.	[27]	[1, 20, 57, 62] [17, 23, 28, 38]	[4, 6, 37, 52]	13
Bias mitigation	Being aware and mitigating bias and discrimination, eg. not using demographic data for decisions.		[23, 28, 56, 62, 66] [1, 16, 55, 64]	[4, 6, 52]	12
Citizen in the loop	Ensuring that citizens know that they are part of the society and their needs are heard, by eg. organizing public debates, dialogues or involving citizens in decision making.	[12, 29, 67]		[1, 6, 28, 62] [4, 11, 13, 30]	11
Benefit and efficiency	Ensuring that the PS AI service brings in efficient manner the benefit not only to the user, but also to the society and possibly the planet.	[31, 32, 67]	[17, 23, 57] [28, 38, 62]	[8, 22]	11
Accessible and good interface	Accessible for every citizen and possibly aesthetic interfaces with predictable navigation and timely feedback about what is happening in the system.	[63, 67, 68] [32, 33]	[23, 56, 62, 69]		10
Privacy	Ensuring the privacy of the citizen, not using their intimate or too old data, storing only needed data and anonymizing.	[19, 23]	[1, 28, 38, 62, 66]	[4, 52]	9
Respecting citizens' rights	Obeying citizens' freedom and constitutional procedures. Being compatible with cultural diversity, social norms and values. Not imposing any lifestyle choices.	[32, 48]	[38, 56, 57, 62]	[37, 54, 60]	9
Data agency and consent	Giving citizens the control of their data. Asking for consent before storing, using and sharing data.	[27]	[17, 38, 56, 57, 62]	[5, 10, 11]	9

Table 3.3: Part one of public sector AI service factors needed for building the trust of citizens.

what	desc	PS	AI	PSAI	no
Education	AI education for citizens via new schools curriculum, public events or courses. Ethics education of people working with AI systems.			[1, 6, 7, 62] [4, 17, 38]	7
Certification	Using a framework or certification to facilitate auditing or ensure quality.	[21, 27]	[1, 6, 17, 20, 38, 66]		8
Accountability	Having people selected as accountable for AI system operations.		[1, 21, 38, 57, 62]	[4, 8, 11]	8
User agency	Users having agency over AI systems, capable to revert or disable it. PS AI services supporting citizens in making better decisions, rather than subverting it.	[30]	[28, 56, 57, 62]		5
Data quality	Using only a data of a good quality: from good sources and not too old.	[1, 20, 28]	[38]	[10, 11]	6
Redress	The redress after AI-made actions should always be easily accessible for citizens, possibly enabling them to talk with human.		[17, 28, 66]		3
Word of mouth	Hearing about experiences of other people and knowing that they are also using the PS AI services.	[67]	[46]		2
Organization	Familiarity and real feel of the organization, by providing timely responses and important organization information, like photo or physical address.	[67]	[19]		2
Experience	Personal experience of using PS AI service		[23]		1

Table 3.4: Part two of public sector AI service factors needed for building the trust of citizens.

mention more general public sector values that are shared between Nordic countries: democracy, openness, service and efficiency. All those are linking to below principles that can be found in different studies and governmental guidelines.

First of all, public sector services should be easy to use and accessible to all citizens [32, 63, 68]. They also should be provided with sufficient privacy. It is also important for the trust of citizens, that services are bringing perceivable, beneficial outcomes [31, 32]. The way that those outcomes are accomplished is possibly even more important. Trust can be negatively affected by any unethical actions, as well as by ineffective or too distant administration. Therein, it is important that the citizens' freedom and procedures based on the constitution are being abode [32, 48]. Last but not least, for building a trust culture it is important to make citizens feel that they are part of the society and their needs are being heard [12, 29].

3.3.2 Trust building factors to AI

In the UK Citizens' jury on trust to Automated Decisions Systems, participants expressed the need to have agency over those systems, in particular, they would like to have a possibility to opt-out from it. Moreover, they voiced that they would trust results better upon them being monitored and assessed by experts [30]. In the study based in Finland, 20 % of participants voted for ethical certification as a way of increasing their trust towards AI, making it second most common need after transparency of the system [21]. Similarly, in the study conducted about personal data usage, 66 % of participants would find a fair data label as an important factor in increasing the trust [27]. The biggest, however, factor was that users should be able to control their data, that is being able to delete it or decline to sell it to a third party. As a solution, 74 % of Finnish participants voted for giving consent to each service provider separately. Other significant factors of building trust were the security and reliability of the service.

In the expert reviews and guidelines, one of the most repeated needs in AI systems is proactive prevention of system failures, therein biases. Most often calls are for proper risk management measures, ensuring safety and reliability of systems [1, 17, 20, 23, 28, 38, 57, 62]. In some sources, the focus is also laid on the importance of ensuring the security of the data [1, 28, 62]. Another important measure for risk prevention is ensuring the quality of data by using only trusted repositories [20, 28]. However, even data of the high quality can contain subconscious cultural biases [1]. Nevertheless, it is important to be aware of potential discrimination and find a proper solution for mitigating it [16, 28, 55, 56]. One of the ways of approaching that is to make sure that decisions are not based on demographic data such as race or sex [23, 62, 66]. Another advice is to obey strictly nondiscrimination and data protection law or to use one of self-regulatory structure such as FATML (Fairness, Accountability and Transparency in Machine Learning) principles [64].

In order to ensure high quality of systems and build trust it is advised to enable their auditing, that is enabling testing and monitoring them [17, 62, 66]. Different frameworks and certifications are suggested for making the audition process easier: standardized metrics for the AI products trustworthiness [17], FactSheets [20], governance frameworks [38], full-stack supply chain [6] or Social Impact Statement for Algorithms [66]. As an example of a highly regulated environment that gained big trust is commercial aircraft. On the other hand, it is important not to hold back scientific progress with law regime [1].

In section 3.2.3, I mention a need to have clear information about who is accountable for the service. Before such a piece of information is published, however, responsible organizations need to select accountable people for system operations [1, 38, 39, 57, 62]. Moreover, it is important to provide an accessible avenue of redress [17, 28, 66]. As a solution, Floridi et al. suggest on following: "An AI ombudsman" to ensure the auditing of allegedly unfair or inequitable uses of AI; A guided process for registering a complaint akin to making a Freedom of Information request; and The development of liability insurance mechanisms, which would be required as an obligatory accompaniment of specific classes of AI offerings in EU and other markets".

Another way to mitigate the risks of automated systems is to monitor it and keep control over it. In details, that could mean having human decide on which decisions should be taken automatically and which by human [17, 38, 57], especially leaving decisions that can affect people's lives to experts [62]. Such a solution is being called human-in-command [28]. Another option would be to have the human-on-the-loop approach, which contains monitoring a system's operations and intervening when needed [17, 28]. Not giving too much autonomy for AI, can decrease risk perception and lead to improving users' attitude to it [3]. Lastly, AI systems should respect and support citizens in making better and more informed choices, rather than subverting it [28, 57, 62].

Looking at the AI systems purely from the user perspective, it is advised that they can be in control over the system and data. They should always be able to invoke needed AI services, but also revert, change or disable them [56, 62], or ask for human interaction instead of automated one [28]. Moreover, users should be also be provided privacy [1, 23, 28, 66]. From the user perspective, they should have control of their own data and be always able to access it [38, 56, 57, 62]. That is connected with the consent term, that is asking users for allowance to use their data beforehand [17]. The other privacy measure is from an organization perspective: it should be restricting the amount and age of data held [38], do not sense data from personal spaces, neither gather any intimate thoughts or emotions and make sure to anonymise personal profiles [62]. Related human-centred data economy, under the name MyData, is currently quickly advancing in Finland and attracting international attention [12].

Important, but often forgotten factor impacting user attitude to the AI tool is its interface. Some of its features can be derived from the transparency and

user control need, as Shneiderman et al. [69] mentioned, they should help users to understand underlying algorithms and give the potential for to better control the tool. Moreover, they cannot imitate human characteristics, like voice or appearance, so it is always possible to distinguish interaction with human and AI [62]. On the other hand, designers should remember that higher anthropomorphism enhances trust in complex systems, as well as aesthetic appearance [23]. Amershi et al. [56] add in their guidelines the need to enable user feedback and keep on adapting and personalizing tool based on users' actions and feedback. Lastly, predictable outcomes and behaviour can also help in growing trust in the system [23].

Another important trustworthiness factor of AI systems is how users are perceiving its benefit and efficiency in a specific task [23]. Different guidelines mention different benefit subjects: well-being of as many people as possible [17, 38, 57], well-being of all sentient beings [62] or societal and environmental well-being [28]. Furthermore, it is important that any AI systems would follow human rights [38], reduce social inequities and maintain bonds of solidarity between people [62]. It also should be compatible with cultural diversity, social norms and values; it cannot impose any lifestyle choices on society [56, 57, 62].

Last but not least, three more single factors were suggested. First, is that the trustworthiness of service can be enhanced by giving it "real world" feel, that is timely responses and information about the organization, like a photo or physical address [19]. Schaefer et al. [23] argues that for many trust builds over time, therein over experiences with technology. Experience, education and mental image influence by media are other significant factors mentioned in the Trust&AI survey [21]. On the other side, Alexander et al. [46] found that the social proof, that is the knowledge that other people are using specific AI system is most effective in convincing others to use it as well. Lastly, in the reviewed literature, many voiced the need for public dialogues and for education, however, those will be discussed in the next subsection 3.3.3 as they merge with Public Sector policies.

3.3.3 Trust building factors to AI in the Public Sector

Similarly as in Transparency section, here too there are points repetitive for the public and private sector. One of the important trustworthiness cues there is a feeling of efficiency of the service and benefit for citizens, rather than for the government [8, 22]. It is also important that the power (or control) balance is shifted more towards the citizens. Talking about personal data, citizens should have full control over its use in AI systems [5, 10, 11]. Specifically, they should be able to choose which of their data can be used, validate them, correct or delete whenever needed. From the organization side, it data should be always evaluated by the quality of its sources, necessity of it, age validity and storage [10, 11].

Furthermore, citizens cannot feel discriminated or fear using AI systems in order to trust them, the safety and privacy need to be provided [4, 52]. The

risks must be mitigated, limitations examined and systems overseen, audited and monitored [4, 6, 37], possibly by multidisciplinary and diverse teams [10]. However, different requirements (e.g. human-in-the-loop, monitoring) should be adjusted to systems of different impact assessed levels [37]. Human rights, democratic values and diversity should always be respected and human intervention enabled where necessary [37, 54, 60]. Systems must not discriminate any citizen [4, 52] and more support is needed for mechanisms and organizations that would help any harmed from AI system citizen [6]. There is also a need for designating accountable people or organizations for AI systems and its actions [4, 8, 11].

The need for education is very often mentioned across different guidelines for trustworthy AI, therein AI systems in the public sector. Most are focused on education of the general public via public awareness activities, expert lectures or bringing easy access to different knowledge sources [4, 7, 17, 38]. The main focus on such activities would be on potential consequences of using AI in society and work, therein possibilities and risks that it brings. That would aim to minimize fear to AI among citizens, empower them and achieve social acceptance of AI. Moreover, it is suggested to promote developing skills such as critical thinking, digital and media literacy [62]. Another suggestion is to bring better AI education to schools and universities [1, 6, 17]. It is both important to educate experts in AI by good technical studies, as well as teach technical AI basics to students in social sciences departments and schools. Moreover, expert education in AI should also include ethical studies, therein possible hazards with using AI. That can help in fostering the ethical AI development and facilitate social acceptance to AI. Lastly, it is important to teach employees inside organizations that use AI, about risks connected with using it [38, 64]. That is an important step in failures prevention.

Proper education ensures the flow from the government, organizations and experts to citizens, however, it is also important to ensure the flow in the other direction. That can be achieved by the "citizen in the loop approach", where the public is invited to development, deployment and oversight of AI systems [30, 62]. Therein, public debates or dialogues are encouraged, where citizens could express their needs and help in creating the limits on AI development and usage [1, 11, 28, 62]. It is also encouraged to support communities that are helping in enabling public participation in AI decisions [6]. Specifically, to the public sector, it is also advised, that any of the AI usages would be started slowly with pilots, that would possibly involve citizens [4, 13]. Such pilots are recommended to be broadly published.

3.4 Guidelines for trustworthy public sector AI services and personas

The primary goal of guidelines is to support designers and developers in their work, however, the ultimate aim is to support future users of developed with

guidelines services [70]. As to my knowledge, there is no study telling what to include in and how to design guidelines for the area of ethical technology. That can be caused by the thesis, that design of guidelines is strongly dependent on their context [33, 71]. Moreover, there is no study yet on what impact using guidelines can have, despite them being used broadly in multiple areas [70].

Indeed, there exist multiple materials on ethical and trustworthy public sector, AI and public sector AI services. There are principles developed by organizations such as The Organisation for Economic Co-operation and Development [54], Finnish Ministry of Finance [68], AI4 People [17], FAT/ML [66], Future of Life Institute [57]; guidelines provided by UK government [60, 61, 63, 65], European High-Level Expert Group on AI [28], IEEE [38] and by two other studies [33, 56]; and finally national directives published by The Government of Canada [37] or for the government of France [72]. The guidelines presented in those documents were used in answering on research questions 2 and 3 in above subsections.

Nevertheless, some recommendations on how to design good guidelines can be inspired from materials on design, usability or clinical guidelines [56, 70, 71, 73]. Firstly, it is advised that before starting guidelines, thorough research should be done by leading discussions and literature review, keeping the multidisciplinary approach [71]. Secondly, the included descriptions should be easily understandable by the future users of it [70]. Thirdly, guidelines should represent clear instructions on how to reach specific principles [56, 73]. Moreover, from the background of this work, we can see that it is advised to include the voice of citizens in the guidelines [17, 30].

An additional tool that can be helpful for designers and developers of the service are personas [74–76]. They describe specific behaviours and attitudes that can be distinguished among people using specific services, help in building understanding between users and creators and hence help in prioritizing service requirements and design [75]. In the best version, personas should be based on the real data about users, as it makes them more realistic and believable [74]. One of the successful cases of using personas is described in the case study led by Randolph [76], where they helped in the design of the small information system.

Chapter 4

Empirical study results

This chapter presents the results from the empirical study of this thesis. The first section 4.1 displays the results from the interviews, the second 4.2 from the design workshop and the third 4.3 from user testing. The last section 4.4 present results summarized in the form of guidelines and personas.

4.1 Interviews

This section displays the results from the interviews done for this study. The first subsection 4.1.1 presents the demographics of participants that took part in the study. The next ones introduce to the results grouped by accordingly attitudes to AI use (4.1.2), needs and concerns (4.1.3) and what information about services would be needed (4.1.4).

4.1.1 Demographics

23 people were interviewed. Two are not counted in, due to poor language skills and being too shortly in Finland, therefore there are 21 people counted in the results. The abbreviations in brackets, like (F), are later used in presenting the results.
Gender: 11 women (F), 10 men (M)
Age: 12 17-30 y.o., avg. 25 (<30); 9 33-67 y.o. avg. 48 (>30)
Education: 12 finished at least bachelor (HE); 9 with no higher education (LE);
Nationality: 12 Finnish; 9 immigrants staying in Finland for 3-20 years (avg. 9 years);
AI Level:
6 of no interest, no knowledge (AI1 - poor awareness)
8 of one of interest or similar background (AI2 - medium awareness)
2 of having interest and some background (AI3 - good awareness)
5 of having strong interest and broad knowledge (AI4 - great awareness)

4.1.2 Attitudes of participants to AI in Public Sector

In different moments of the interview, interviewees were often expressing their attitude towards Artificial Intelligence used in the public sector. The attitude could be divided into their feelings, concerns and way of acting with services. During interviews mainly first two factors were identified and they are depicted in this subsection and the following subsection, first feelings, next concerns. Both topics are firstly presented in the summary from the whole interview, later divided into feelings raised in specific interview moments.

To understand the general feeling of citizens towards using AI in the public sector, I categorized all mentions of feelings from the whole interviews to three categories: fine, mixed and negative. Later, I put the aggregated feelings from all interviews to the chart 4.1. As seen, there were more positive feelings mentioned when talking about using AI in PS.

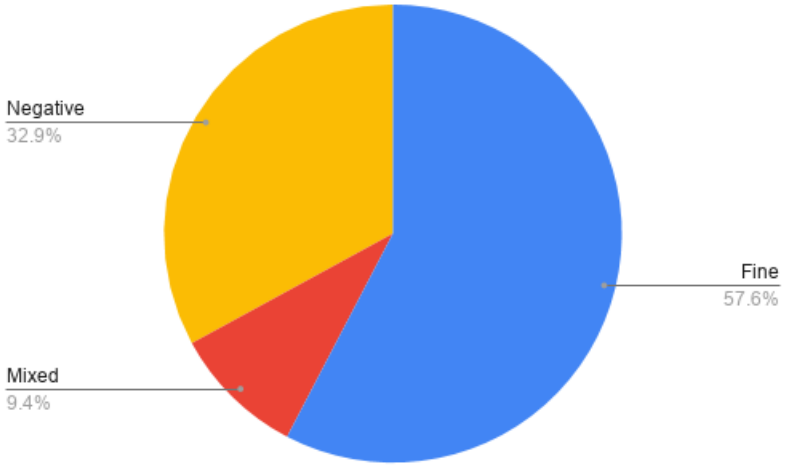


Figure 4.1: Aggregated feelings towards AI used in the public sector from interviews.

When looking at the demographic comparison in the table 4.1, we see that there is a significantly higher number of positive feelings mentioned in a group of female (3:2), high educated (4:3) and immigrants (3:2). When looking into more detailed groups based on their AI awareness, educations level and age in the table 4.2, we see different dependencies. The most negative or concerned attitudes towards using AI in the Public Sector can be seen in the group of highly educated and specialized in AI men around 30 y.o (1:2). Same educated group, but older (avg. age 45) is definitely more positive towards AI(5:1). Another group of people who are highly educated, but with lower AI awareness, around 27 y.o. is visibly more positive about the topic (2:1).

Table 4.3 represents distribution of feelings mentions per different interview parts. The interview started with asking for an opinion on citizens personal data being used in the Public Sector. Nine participants, most of which of lower AI awareness, said to be fine with the use of data in the public sector.

	AI1	AI2	AI3	AI4	F	M	HE	LE	Fin	Imm
neg	6	16	5	15	19	23	27	15	27	15
pos	13	20	5	15	30	23	36	17	26	27

Table 4.1: Demographic distribution of positive and negative mentions of feelings towards AI. Legend: neg - negative, pos - positive, AI1-4 - AI awareness levels from lowest to highest, F - female, M - male, H/LE - High / Low education, Fin - Finnish, Imm - immigrants.

	AI34 HE 30	AI34 HE 45	HE AI12 27	LE AI12 22	LE AI12 55
neg	18	3	7	7	8
pos	10	15	16	11	6

Table 4.2: Distribution of positive and negative mentions of feelings towards AI across different groups of participants. The groups naming is created from following parts: AI12 and AI34 mean people of low (AI12) and high (AI34) AI awareness, HE and LE mean people of low (LE) and high (HE) education status, 30, 45, 27, 22 and 55 are representing the average age of people in the group.

Part of them explicitly said that they are trustful: ”you’d be interested only if you are doing negative things”. On the other hand, four participants, most of which were immigrants mentioned that they are suspicious or worried about the use of data in PS: ”I want to believe for good reasons”. Mixed feelings were generated by the group of great awareness of AI. All of them admitted to be fully trusting PS, however, there was always a ”but” in their answers followed by worries and suspicions of possible failures. They use words such as ”comfortable” ”call me naive” to describe their feelings.

For comparison, similar questions were asked about using data and AI in the private sector. There, most of the people of a good awareness of AI were very negative. Words like ”helpless”, ”necessary evil” or ”hate it” were mentioned. Only one person from the group mentioned that is not scared, another, of medium AI awareness, call themselves interested and yet another of a poor AI awareness, that is not bothered until it goes to wrong hands. Five low educated people admitted that they feel like they are being observed. Another five participants of higher education claimed to be active towards it by reviewing terms or modifying data sharing options. Three younger participants said not to be interested and trying not to think too much of it. One also said that AI makes you: ”feel less of a person more like a customer”

Furthermore, citizens were asked about their attitude solely to Artificial Intelligence. There, voices were very different: from being fairly keen on it to not interested at all. A few participants who had longer experienced with working with AI seen it as ”nothing magic” or that is already everywhere: ”we cannot stop the train anymore”. Two lower educated participants mentioned that they are scared of AI. One other said:

> I don't think we can stop these (AI) from coming or influence things and make these more human. Everything is happening too fast.

	GK	UC1 (13)	UC2 (17)	UC3 (16)	UC4 (14)	FU	total
fine	8	8	11	6	6	10	49
mixed		1	2	3	2		8
negative	4	4	4	7	6	3	28

Table 4.3: Feelings voiced in different parts of the interview: GK - general knowledge and attitude, UC1-4 - use cases 1 to 4, FU - Follow Up questions. The numbers in the brackets by UC1-4 indicate how many times specific use case was presented during interviews.

In the next part of the interview, each participant was asked to give opinions on two or three use cases. The use cases were given in different orders to avoid bias and missing answers to some of the cases. Therefore each case was mentioned between 13 and 17 times.

The first case represented situation, where the online application to public sector service is pre-filled with the personal data coming from either government's database or from other public sector organizations. 8 people viewing the case were positive about it. Five of them explicitly liked the automated decision process, reasoning that it makes the process faster and incorruptible. 5 of young and highly educated ones admired specifically autofill, as a way of saving time. Moreover, two mentioned that there are happy about the data flowing between different PS organizations. On the opposite side, four people were negative about the case, three of those having low AI awareness and low education. They saw the case as harsh and non-human process and did not like having any influence over it.

In the second use case, AI was used to predict and inform about possible social health problems of relatives. There, 11 asked, of average medium AI awareness, felt positive about it. They liked that with such a service, they could help their relatives. Two other people felt split: "I feel mixed emotions. [...] seems like, they might take too much information" and two other uncomfortable: "It would be strange for me to have information about other's illness. Scary, I would not want to know.", "it should be social care taking care of it". Finally, for two of asked the level of information that would need to be gathered for such a prediction overweight positive effects of the case, thus they expressed a negative attitude towards the case.

The third use case depicted the education impact on kids assessment, by collecting the data from smart devices used by kids. Only this case gathered more negative (seven) than positive (six) opinions, while three interviewees weren't sure about their feelings. Worth to mention, that it was mostly people of low AI awareness who did not like the case, feeling that it would be too much of surveillance: "The first feeling is that now the border was broken. Following my child 24/7. It's disgusting to think", "(...) it sounds so weird.

Like being a test rabbit.". Out of positive voices, four were coming from highly educated people.

Last, the fourth case presented the use of AI for discovering frauds connected to public benefits. Here, the feelings were equal. Six recipients, mostly highly educated and under 30 years old, felt bad about it. As reasons they stated being uncomfortable with data coming from unknown sources, having no human in process and with the message being not enough transparent. On the contrary, six people, mostly lower educated, were fine with such a case. They saw it as the right thing to do.

At the end of the interview, interviewees were asked a few follow up questions. Some of them were directly asking about their attitude for AI in the public sector, after seeing interview cases. In an answer, 10 expressed good feelings about it: "It is good that artificial intelligence is being used and people are becoming conscious.", "I see a lot of benefits in these, for example in terms of the child or social welfare.". Five people mentioned that it is inevitable to come anyway. Three people of moderate AI awareness mentioned negative emotions like being worried or feeling weird.

In summary, talking specifically about the trust, participants admitted having more trust in the public sector handle their data than to the private sector. As a reason, some mentioned: rules and regulations PS is obligated to abide, its righteous history and no commercial interest. Nevertheless, people seem to have even less knowledge of how data is used in the PS compared to the private sector.

A few people, who have bigger awareness also noted a difference between the trust to people handling the service and the AI itself: "I don't really have mistrust towards the AI. But I do have mistrust towards people.". What is more trustable, however, would be debatable, as people can be biased and corruptible, while AI can be erroneous: "human might in some ways be worse, because of the human error. [...] AI can do the same mistakes, because of humans fed the data", "humans are biased. I think that's okay, systems can also be biased, but at least they're fairly biased. But humans are biased in ways that are very unpredictable.", "Because if machine is really accurate, then it would be fine. But [...] machine can also make some mistakes.".

4.1.3 Concerns about and needs for public sector AI services voiced across the interview

Similarly as in the feelings, different needs and concerns were stimulated by different interview parts. In general, the charts 4.2 and 4.3 shows how many times different needs and concerns were mentioned by all interviewees during the whole interview. Following tables 4.4 and 4.5 presents how much were those needs and concerns repeated in specific parts of the interview. Specifically, interview-part-related ones are described more in the below paragraphs.

Looking at the demographics, the most concerns were brought up by the group of AI specialists, while the least by a group of young, low educated par-

ticipants. Specifically, surveillance was mostly mentioned amongst the former group, while security was for the older high educated group. As for the needs, the younger the participants were, statistically more of those they mentioned. Especially the need for consent was raised very often by the younger than 30 years old group.

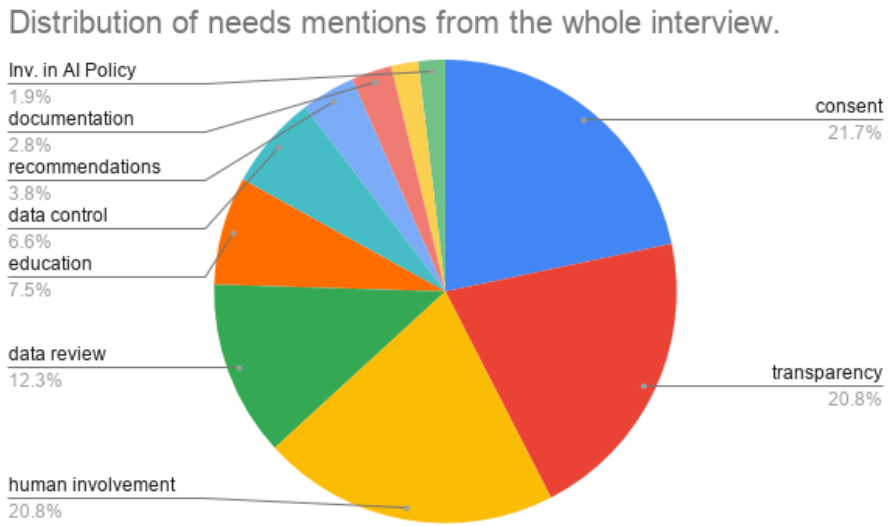


Figure 4.2: Distribution of mentions of needs from the whole interview.

	GK	UC1 (13)	UC2 (17)	UC3 (16)	UC4 (14)	FU	total
consent	7	1	6	4	0	5	23
transparency	0	3	0	4	2	13	22
human involvement	0	4	4	2	3	9	22
data review	0	2	0	2	3	6	13
education	4	0	0	0	0	4	8
data control	0	0	0	0	0	7	7
recommendations	0	0	3	0	1	0	4
documentation	0	3	0	0	0	0	3
updates	0	0	0	2	0	0	2
Inv. in AI Policy	0	0	0	0	0	2	2

Table 4.4: Needs voiced in different parts of the interview: GK - general knowledge and attitude, UC1-4 - use cases 1 to 4, FU - Follow Up questions. The numbers in the brackets by UC1-4 indicate how many times specific use case was presented during interviews.

At the start of the interviews, the most commonly repeated need was about the consent for using data and that it is very important to trust PS: "It's ok for the public sector to use and share my data as long as I can track the use if I want to.". Those were mostly people under 30 years old. Four other people

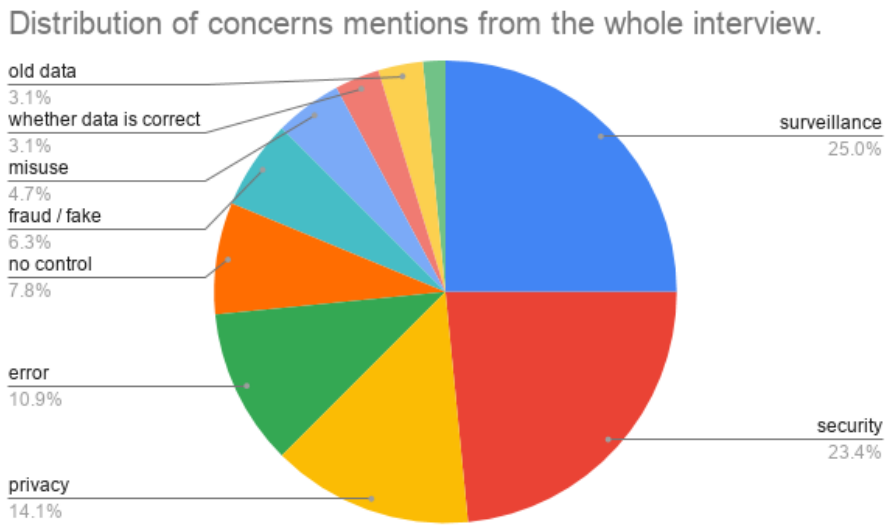


Figure 4.3: Distribution of mentions of needs from the whole interview.

	GK	UC1 (13)	UC2 (17)	UC3 (16)	UC4 (14)	FU	total
surveillance	0	0	6	4	6	0	16
security	3	3	0	3	0	6	15
privacy	0	0	2	5	2	0	9
error	0	2	0	0	0	5	7
no control	0	0	0	2	0	3	5
fraud	0	0	2	0	2	0	4
impact	0	0	0	4	0	0	4
wrong data	0	0	2	0	0	2	4
misuse	3	0	0	0	0	0	3
scoring	0	0	0	0	0	1	1

Table 4.5: Concerns voiced in different parts of the interview: GK - general knowledge and attitude, UC1-4 - use cases 1 to 4, FU - Follow Up questions. The numbers in the brackets by UC1-4 indicate how many times specific use case was presented during interviews.

mentioned the need for education for themselves or others. Only concerns that were heard then were about the security of data, for example, due to bad technology providers, and about the misuse of data. The latter one was, however, only mentioned by immigrants. Moreover, the doubt whether PS is capable of managing AI was mentioned:

> The only thing that worries me in the public sector is that do we have the best people to keep the data protected? That worries me the most. [...] and then Google, Microsoft, and Facebook, they are the best in protecting their day to day basis business, they make money out of it. So they had their motivations.

In the first case, the mainly repeated need was for having a human involved in the decision: "I'd rather be treated by someone. I think there would be a better chance of getting an apartment then". Next, it was for the proper documentation of the process: "the first question that arises: is this documented properly?". A few participants, more knowledgeable in AI topic, suggested that it would be interesting to see also algorithms or code used there, however more because of their interest rather than to improve their trust. The other need was to be able to check the data that was used for the decision, whether it is not, for example, outdated. Moreover, two people added a comment that such data sharing and auto-filling is ok in the public sector, but would never be acceptable in the private sector.

In the second case, the most frequently repeated needs were for the consent of the tracked person: "I think they should first ask about the use of the data." and for being contacted by a person in such sensitive case: "I'd rather have someone tell me directly than email.". The most common concern, mentioned mostly by highly educated and AI aware ones, was about surveillance: "There is thin borderline before surveillance society. So it can provide a lot of good things, but then the next philosophical question is, who is watching the guards.". Next, several people were feeling that it would be a privacy violation, fraud or that the data are wrong.

For the third use case, again consent was in top-repeated needs. This time, however, it was also specifically mentioned that kids should be able to agree to the following tracking. Other equally often raised requests were for the transparency: "If I'm convinced that they get the results with the use of these two devices, then I can further take how bad it is. Then I can decide." and regular updates of the process. The case raised plenty of concerns, mainly about privacy (whether children location or messages will also be tracked) and impact on children (mentioned mainly by people above 30 y.o.). Moreover, some did not like that children would be constantly tracked or that they would have little control over the process.

In the fourth, fraud detection case, the needs were for the consent, clear rules and laws that enable for such actions and having human involvement in the process. The biggest concerns were about being under surveillance: "Just because it is like clear oversight, they have to be there monitoring, you know, directly.", "There seems to be no privacy" and data being shared between different sectors: "Identifying cash would be a little too much. Then there is too much data in use.". Those concerns were mostly mentioned by young native-Fins of moderate AI knowledge.

During the follow-up questions, the need for transparency was raised thirteen times: " It is very important that you see what information is and can affect. And young people and children should be told about these. Maybe classes will be held in schools and for seniors". Half fewer mentions were for being in control over data and service: "we should be in more control of switching it on and switching it off what we do and do not share." or at least for being able to review it: "I think I should have the right to review the data."

Similarly to the first part, education: ”Maybe I should read some articles myself. And Kela could provide more information and links to relevant articles”, ”I would like to see videos and learn what this all means. Visualizations, no long writings.”; and consent were also brought up: ”You should always be asked for permission to share [data] and would not automatically pass on.”. Two participants also mentioned that they would be happy to participate in the sessions for creating PS AI policy: ”I think most of these are in setting the policy of what is okay and what is not. I think that the user can have a voice into how they feel about things in general, in the same way how we vote to pass legislators and pass laws like this is the set of information and it is okay to use and this deceptive information that is not okay to use.”

Two of the most commonly repeated concerns were security and privacy, mostly by over 30 years old, experienced highly educated and in AI part of recipients: ”Security is really important in a lot of these, isn't it?”. Mainly, they worried that someone could steal a citizen's identity if public sector has a common database with all data of their citizens, which would be badly protected: ”So I think I think it is maybe that all the government sectors have a shared database. [...] maybe it is too much information in one place. And if it gets leaked, then people could full falsify your identity easily with all this data”. Other worries were that the public sector does not have itself best-experienced people to take care of data safety, neither the public sector technology providers are best in this area: ”I am suspicious because I know that the public sector contracts the private sector, for building applications, [...] And every time you read about it, most companies just do a shitty job securing their data when they just want to push out the product as possible. [...] So there's this gap between experience like, well-done software design versus what the public sector is able to pay for”. The worries were added about AI is not yet accurate enough to use in such cases and about using old and no longer relevant data:

> But at the same time, I also think that so many things happen over the course of a person's lifetime, that, in turn, enters a dangerous territory where the data follows them around. I feel people should be able to use more up to date data, or shed off certain parts, for example, if they had financial troubles in the past, in no way should that impact their financial situation 10 years later, because so much can happen in 10 years. So it needs to be done in a way that's a bit ethical in the sense that, you don't carry your data around with you, from the whole lifetime, up until that point that you should be able to phase out certain parts to like, rebuild your reputation.”

Regarding, where human should be in the process, there were several approaches across the interview. Firstly, many recipients mentioned that interactions such as asking for consent, explaining the process, giving the results and negotiating should be handled by human: ”[...] we can replace as many

things as possible with machines. But at the end of the day, we still crave human interaction in some form or another. So I don't think we'll ever be able to fully replace the human in this scenario". The more personal the case (e.g. the health prediction) the more it is important for the information being passed via a person. When it should be the decision to pass, that is based on the specific rules, it is more acceptable to get the information automatically: "AI can't be trusted as much as humans. Data is in safer hands with humans. AI should be used for low-level tasks only". However, some contact to discuss it should be provided. Secondly, part of interviewees preferred human to make decisions, especially in critical situations. That means either human using AI as an assisting tool or reviewing the data and the decision before accepting it: "I guess when the decision is very important, like, whether someone goes to prison for a long time, or if someone has to take some dangerous drugs [...] the computer should be an assistant for a human, but not the deciding component". Lastly, some mentioned that a person should be monitoring the whole process: "Nothing should be automated, when it comes to analysis and evaluation, you have to have someone who can verify that the system is working according to rules and ethical guidelines, as demanded by society".

4.1.4 Information about public sector AI services requested by citizens

As presented in table 4.4, one of the most important needs for citizens is transparency. To answer a question of what transparency means for citizens, we presented cases during the interview that were very scarce in any information about the service. That was intended to stimulate participants to express what information about the service they miss. In the table 4.6 a reader can find what type of information were requested by interviewees in different use cases, sorted by the total number of mentions.

Explanation of a specific decision was the most repeated need across the use cases, especially in the first use case, where applicants were automatically rejected from receiving housing: "[... I would] have to call somebody to try to figure out why. Then they also have to figure out why. So, if it is smart enough to decide immediately why I'm not going to get the house, it should also be smart enough to tell me immediately why". There, specific criteria for a decision were also requested. In last (discovery of fraud) and the second one (predicting illness of a relative) it was mentioned as well: "I would like to know what has happened, on which basis there is reason to doubt". Moreover, the fourth use case stimulated some responders to ask for specific rules and laws behind the service.

The next top questions were connected to the data: where does the service get the data from and what kind, and how much, of data they have: "We should be told everything about research and personal data use. It's suspicious if the information is not given. People get paranoid". The data source was especially important in the fourth use case, where interviewees were wor-

	UC1 (13)	UC2 (17)	UC3 (16)	UC4 (14)	total
Explanation	10	4	*	4	18
Data source	5	2	*	8	15
Data used	*	4	6	5	15
Process	1	5	1	6	13
Purpose	2	2	6	7	13
Impact	0	0	4	0	4
Relevance	0	1	2	2	5
Data storage	0	0	4	0	4
Regulations	0	0	0	3	3
Service stakeholders	1	0	2	0	3
All personal data stored	0	2	0	0	2
Redress channel	1	0	0	0	1
Who can access data	0	1	0	0	1
Accuracy of results	0	1	0	0	1

Table 4.6: Information about the service that were mentioned in different parts of the interview: UC1-4 - use cases 1 to 4. The numbers in the brackets by UC1-4 indicate how many times specific use case was presented during interviews. The star indicates that the information was included in the material or non-relevance, thus there were no question for it.

ried that information might flow from private to the public sector. Related to data, some people also asked for specifically all personal data service might have and who has access to the data. People with IT background were also interested in how and how long is it stored.

Next two pieces of information needed are process and the purpose of using AI in the service: "So there should be transparency about purpose. So what is the intended purpose? What is the reason this service exists?". For the former, interviewees were generally interested in easy words how the data is collected, used and processed and whether there were people involved. Only one person of high AI awareness additionally asked for the accuracy: "And then the performance? How confident are they?". For the latter, the purpose was requested mainly in the third use case (education impact analysis), while no question about it was asked in the second one. Related to the topic, some individuals also asked for the reason of using specific data, why does it matter and benefit to the informed person and what's the interest of the stakeholders of having it:

> There should be transparency about purpose. What is the intended purpose? What is the reason this service exists? The process? How do they do it? How did they use your data? And what kind of conclusions are they trying to get out of it? And the performance? So the results how confident are they?

Other questions that were mentioned were about the impact on the users,

point of contact to make a complaint or inquiry, service stakeholders and the owner. Regarding the former, that was mentioned only in the third use case, where interviewees were worried about the impact of the service process and results on the children. Two last ones were often mentioned in the messages, however, still some asked for specific information about those.

4.2 Design Workshop Results

In the below sections, a reader can find results from the design workshop. During the workshop, people were asked to note down or draw what they would like to see in 3 suggested services, that contained one of decision making, prediction and impact assessment, all automated with AI. Results start with demographic 4.2.1, after which results from each service are presented in sections 4.2.2, 4.2.3 and 4.2.4. All of those are summarized in subsection 4.2.5. In the service-related subsections results are grouped by the touch-points with the customer (e.g. before use), while in the summary, they are grouped by topics (e.g. consent, data).

As participants were not restricted to one type of discussion presentation, there were multiple different types of materials created, therein: post-its notes with answers on questions; loose notes and answers written on the plain paper; an interface design or a message design. During the analysis process, all of those were used in the same form or were rewritten to post-its.

4.2.1 Demographics

For the design workshop there were eight people were invited, four women and four men. Age of participants was between 22 and 38, where the average was 28. All of the participants finished at least the bachelor level of studies. Three of the participants were born in Finland, the other five stayed in this country for an average 6.5 years. One person at the workshop had poor knowledge of AI (AI1), three people had medium awareness (AI2), two had good awareness (AI3), and two were working in the field of AI (AI4), so the average in the 1-4 scale was 2.6.

4.2.2 Transparency and other factors needed in decision making AI case

In the Decision-Making AI Case participants were deciding on what and when to inform the users as follows. Before the decision is made, it is suggested that the user should see:

- all the data that is submitted and what would be considered to the decision making
- where is data coming from (organisation name and to what other data they have access)

- where is the data stored and/or is it safe and secure there?
- what is the purpose of using specified data?
- how will the data be used?

After the decision is set, the following information was suggested to be visible:

- decision,
- main factors and/or data used for the decision,
- reasoning on how the decision was made,
- clear steps on how to inquire or complain about the decision,
- where to get more information about the service,
- the principles of the specific process,
- visualisation showing demographics of rejected and accepted applications (for understanding if the system is any biased),
- if the data is shared with any other organisation (especially if with the private sector),
- if data are public or private,
- if the data was used for something else,
- process of the service.

Moreover, participants mentioned that all of the information should be simple and visually comprehensible so that there is an easy way of understanding it. There should not exist information overload, as that would be confusing. Visualisations were suggested as one way of achieving good results and information representation. It was suggested that the application and receiving the message should be done digitally through online service. The final decision should be easily accessible also any time after using the service. Alternatively, it was also suggested that some decisions could come via phone calls.

While discussing the solution with the whole group, participants mentioned some of their concerns towards the case. One of the started topics was about positive discrimination, that means the example situation where one who applies would not get the flat because the system would choose people who are more in the need. The worry was, whether such applicant would still support such a decision. Another concern was about sharing and storing data. People worried how secure would it be and whether it would not be shared with a third party afterwards. Last big topic mentioned was about the AI bias, even though at the beginning 3 people knew only what it is.

For rising the trust of the service, participants suggested that users should always be able to resign from sharing their data with other services and from using AI for making the decisions. For the latter, the suggested solution was to be able to ask for or even pay more for human handling the case instead of an automatic process. Furthermore, it was said that human should oversight the decision and that it should be possible to review and update the information given. The additional trust would bring consent for using their data, information that the data sources are reliable and that their data is not shared with the private sector. The service should also have an easy way of appealing and inquiry for more details. Lastly, the following was mentioned: "[we must be able] to trust that the data is being used for that purpose being told".

Regarding the bias, participants suggested that it should be definitely tackled by the service owners, as well as owners should be transparent about it. For example, they could mention in some article about the biases that are right now discovered in the service and write what they do to handle them.

4.2.3 Transparency and other factors needed in predictions by AI case

In the case of predictions made by AI, some groups designed the full message (Fig. 4.4) of what the user should receive and others proposed an app interface (Fig. 4.5). The text of the full message was proposed like this: "Using XYZ data, based on analysis of this data, your risk of getting X disease is higher than the threshold level. Further information by this link: ....", while in the page accessible by the link it was suggested to include information about what data exactly was used, how to access it and who else has the access, whether the process was accomplished by human or fully automated, whether it was public or private.

Apart from above, a few other information to be shown in such service were mentioned:

- Advice, for example contact with the medical expert who could explain the situation or any other recommendations on what to do with the information.
- Data, mainly what was used for the prediction, what was the source of the data and who has access to it. Seeing data also allows the user to understand the reasoning better and see whether all information used are correct.
- Process, so what was the role of AI and human and what system was used.
- Performance, so stating clearly that "This is not a diagnosis and does not replace medical professionals" and encouraging people to not taking the prediction as something 100% sure and giving the level of accuracy of AI.

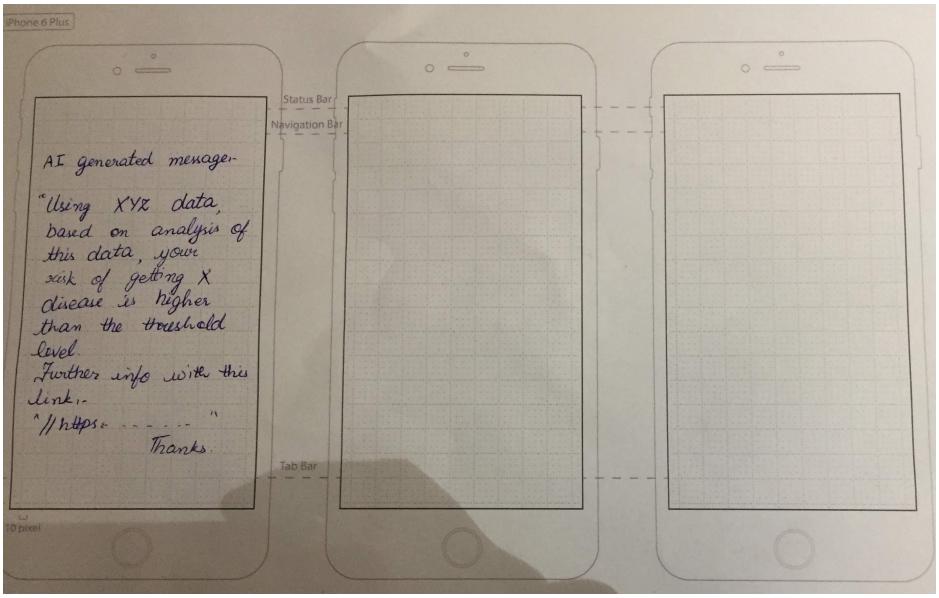


Figure 4.4: A suggestion of the message received in the Public Sector service predicting future diseases created by one group on the Design Workshop.

- Giving access to more detailed information.
- What was the reasoning behind the results.
- The purpose of the service.
- Where and when the user consented to the service.

It was suggested, that the first message should reach the user by the human personnel, email or post. The message should be as clear as possible. More information should be only available in the secured mobile application or website, alternatively by call or meeting with a medical or service expert. Part of participants was arguing that the first or the biggest interaction should be handled by a human, as the data is personal and might affect the user. Furthermore, one should always agree on having such a service, possibly also choose when they would like to be given predictions (e.g. on request or when prediction confidence is above a specific level). Lastly, you should have an option of quitting the service.

4.2.4 Transparency and other factors needed in impact assessment by AI case

In the third case, participants were designing the interaction with the service, where the impact (e.g. on education, well-being or economy) is assessed by AI. Three touch-points were pointed out of the service with people, whose data would be used. Before the start of the assessment, involved people should be informed about the service and asked for permission to use their data. It

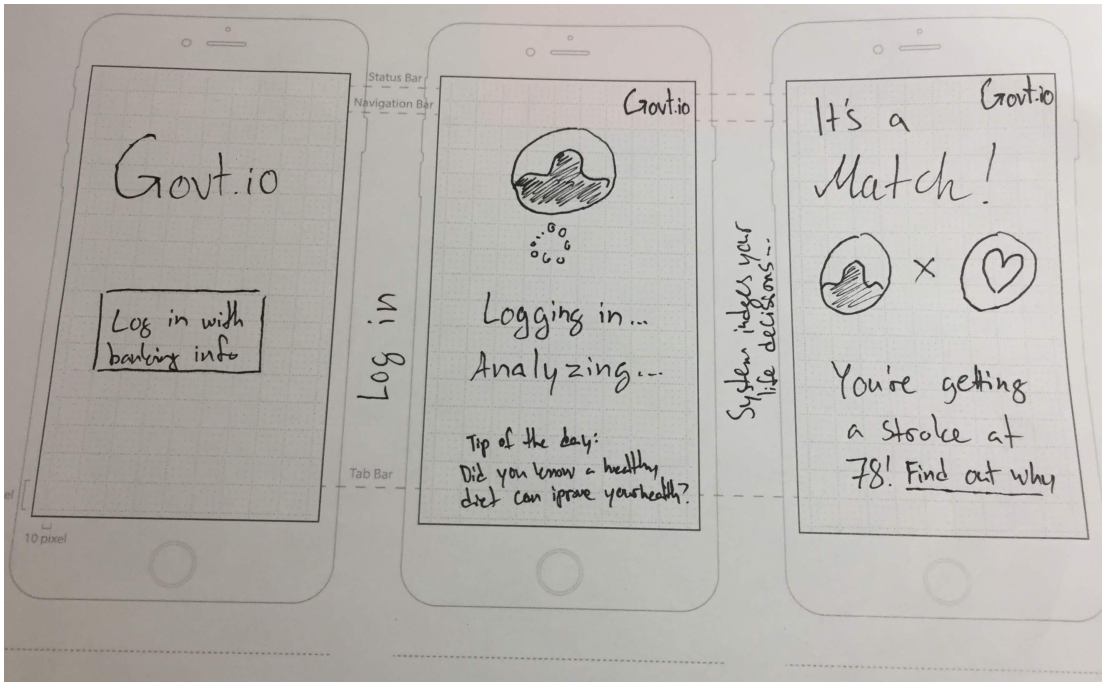


Figure 4.5: An interface of the Public Sector service predicting future diseases designed by one of the workshop participants.

should be known, who will have access to it. During the process, those involved should be informed about stages of the assessment program and have the option to quit it, eg: ”Here are the results, would you like to continue in the data gathering process?”. After the project is finished, data providers should be informed about the results, future privacy of their data. It would be good to present the results in such a way, that makes people who gave the data understand their role in the process: ”based on the results you can predict certain factors. Do you want to play with it?”.

Most important information should be handled by a letter or email, with the indication on how to access e.g. portal with all other information. The ”before” message should be sent early enough, to give time for a decision to possibly involved people. Two groups stated that people should own their own data and be able to give permission to use specific data sources and data pieces. Moreover, they should always see what data is used. There was one discussion about whether involved people should be informed about their data being used if it is used fully anonymously. Two groups were confident that the person always should be informed, while the third stated that it might create bias in results, as a person who knows that their data is being collected, might behave abnormally.

In summary, such information was requested to be seen in the service:

- who has access to the data and results,
- where will the results be published,

- what will happen with data after the study,
- data sources,
- what data is used,
- the purpose of getting data, what will it be used for,
- governmental policy on how the data will be processed,
- contact details to a person who can give more information,
- direct access to the project process (e.g. to the web portal).

Additionally, groups discussed the difference between public and private sector using such data. While it would be ok to use it in the public sector for improvement of public services or research, using AI and personal data to capitalising in the private sector was not well seen. It was also said, that data should be always anonymized. Moreover, if the group of data providing people is too small, it might be still possible to identify specific data, while it is very vital to make sure, that no data can be identified.

4.2.5 Grouped requests for transparency and other factors

After analysing workshop results separately for each case, they were all gathered together and clustered based on topics. In the first cluster, I gathered notes from participants from each group about **transparency**, as shown below:

- ”What information about me will be sent through/considered”
- ”Ability to look under the hood in an easy way”
- ”based on what data was the decision made and reasoning on how was it made”
- p.p.p. - performance, process, purpose
- full disclosure
- as much info as possible + direct access to the project progress

In the next cluster about **process** it the most often mentioned information was the reasoning for a specific decision, prediction or action. Next, in two cases, the main factors for making a decision, prediction or assessment were requested. The single time mentioned was principles used in the system, who is behind the service, what was the role of AI and more information about the system.

Definitely, most results gathered the **data** cluster. The first more important request was for all the data used for decision, assessment or prediction. One

additional voice was for seeing all the data that the service has about the person. Next, often mentioned, was the data source - where was the data collected from, if applicable, what is the organization name and what other data do they have. It was also mentioned that data sources should be reliable. Safety was another big topic. It should be provided for stored data to remove the risk of data leaking or identification. Users should have the information and assurance, as well as they should be informed about what will happen to their data after the use. Lastly, many concerns were mentioned, especially about sharing data with the private sector. Many times it was requested to know who has access to data and whether it will be shared. Moreover, some asked for access to the data.

The next big topic was about **consent**. Participants asked to be able to review the used data and to update it. They want to be able to choose which data and data sources can be used and to always be able to quit the service. If in the service the data is shared with other organization or services, users should always be able to reject such sharing. In general, they want to have control over their data use and their own participation in different services.

The following cluster is all about the **role of the human** in the process. The most often mentioned requests were for being able to communicate with a person: both to get the information about the system, but also to appeal after the results are published. Also, it was twice mentioned, that it should be possible for service users to ask for human oversight of the process. Optionally, it was also suggested, that a person could quit AI use completely. Lastly, participants voiced their interest in general about what was the role of human and AI in the process of decision, prediction or assessment making.

At the **last contact** from the service, such information was requested by participants. Firstly, they asked for a contact to a person from whom they could learn more or appeal, alternatively, a link to a place with more information could be provided. Next, the results should be clearly and understandably presented. Including recommendation or even clear steps of what to do after was also suggested. With the results, there should come both the level of confidence as well as the general accuracy of the system. Consumers of the service should be informed, that the prediction, decision or assessment might be erroneous.

Last but not least, about how **the message** should be designed and transferred gathered following points. Firstly, it was emphasized that all the information should be easily accessible, and the message should be clear. It was suggested that some visualisations can be used for help. The most groups were suggesting to transfer the information via web portals. Some also suggested app or being contacted via emails, phone calls or letter. The message should be always given as soon as possible, early enough to have time for consideration, before taking any next steps.

Regarding the trust, participants played a small game at the beginning of the workshop. Firstly they stood in the line showing their awareness of AI, from lowest to highest, next they were to create line showing their trust to AI,

from lowest to highest. Interestingly, the line reverted: most of the people of the highest AI awareness went to the beginning of the trust line (lowest trust level) and vice versa.

4.3 User Testing Results

For the user testing, external partner created a web-based service prototype, as described in section 2.2.4 and presented in appendix E. The prototype was designed based on the guidelines basing on needs expressed in previous interactions with citizens. In the fake service, coming from the public sector, customers were offered predictions of their possible health issues in the future. At the beginning it was divided into three stages: consent, where clients could choose which data they want to share; during-phase when the data was being processed; and results with the possibility of sharing the results with other organizations.

The user testing was done in three iterations: pilot (not used in results), first and second. In the following sections, I start demographics in subsection 4.3.1, followed with results specific to the iteration in subsections 4.3.2 and 4.3.3. The section finishes with presenting grouped results that repeated irrespectively of the iteration in subsection 4.3.4.

4.3.1 Demographics

Twelve people participated in the user testing, four in the pilot testing that is not counted into below results, four in the first round and four in the second round. Among the latter eight participants, three were female and five were male. Three had a high school education and five higher education. One person had a great awareness of AI, two good one, three medium one and two poor ones. Four participants were under thirty years old, with an average of twenty-two; four others were thirty years old or more, with an average of fifty-two.

4.3.2 Round 1 findings

In general, the service was considered easy to use with a proper amount of text. However, some information about it was pointed out to be missing. Two testers were feeling that for the data that they are sharing with the service, they feel they receive too little - both in results and in general information about the service. Despite service flaws, testers trusted the service enough to share their information, mostly due to the trust in the Finnish public sector. Regarding the interface itself, only one person had trouble immediately understanding how to use the service. This might be caused by the changes appeared with automatic translation of the website.

The information that was missing from the consent page were:

- how was the data collected, data source
- purpose of the service, why is it done this way
- reason for sharing specific data
- example of the data
- clear explanation of the AI processes behind the system
- information about the service process, what are the steps, how long they are
- what happens if a customer does not agree on sharing on some or all data.

Moreover, one person suggested having an open field that would enable customers to write comments about shared data, they said: "Not good that it is only online, human interaction would be preferred. I remember doing a 40-year health-check online and that was an invitation from Helsinki City. There were only multiple choices, no free text option to give personal input and no human involvement. There were no real recommendations and it was disappointing". Furthermore, the privacy statement should be presented in the very beginning, as it answers on questions that might appear while going through data consent: "Had I known that the detailed data is not shown to a human just the results, then I would have trusted to give a lot of data".

The information missing in the results section:

- more details in general
- explanation of the prediction, what factors were taken into consideration
- contact with a person to talk about those results
- guidance about how to understand the results, e.g. percentage of likelihood of illness appearing could be not fully understood
- confidence (accuracy) of AI used
- whether results would be shared with other organisations anonymously or not
- what happens if a customer does not agree on sharing on some or all data

Furthermore, 2 people would prefer having more information, e.g. about the AI process or fighting against bias easily accessible from any point of the service. From the interface perspective, the "all data is now removed" at the end of the results page felt strange that it is already removed when reading the output.

4.3.3 Round 2 findings

Before the second round of the user testing, the following changes were implemented:

- the prediction accuracy was added to the results page,
- privacy statement about how the data is used was moved up in the consent page,
- reason for giving the consent for specific data was added,
- information about how long will the data be stored was added,
- the introductory page with more information was added,
- subtle interface changes were done for increased usability and readability.

Most of the implemented changes were met positively. Again, the message that the service has a good level of information was heard. The first page with more information was liked and suggested to be available from any point of the service. The likelihood of specific health problems presented as risk-levels (low, medium, high) was well adopted. The sentence about removing the data in the last service page was still not understood.

This time, following requests for information, were mentioned:

- risks or impact of consenting for data use, using service, getting and sharing results,
- reason for sharing results with other organisations and whether those are shared anonymously,
- what the predictions are produced in the process,
- time period the data collected,
- sensitivity-levels of the data collected,
- who is the service provider, where it comes from and who is responsible for the service.

Moreover, one person suggested adding explanation that data will not be shared with anyone without the user's agreement already at the consent stage.

4.3.4 Gathered results from two rounds

Below we present what information was still missed in our prototype, that participants asked throughout both user testing rounds. The number by the item indicates how many people mentioned specific information to be missing. In the consent part of the prototype, those were information missing:

- (4) well-explained process of the service, therein the roles of AI and human in it,
- (4) purpose, therein who is receiving the value from the service,
- (2) data period of data collection,
- (2) easy-to-understand description of Artificial Intelligence processes,
- (2) risks and impact of using the service,
- (1) who is the service provider,
- (1) who is responsible for the service.

Moreover, one person suggested having a data sensitivity marks, that would help citizens in estimation which of the data is more or less risky to share. It was also recommended not to use too old data.

In the results part of the prototype, those were concerns voiced:

- results can cause anxiety in the user, especially when they show personal risks and issues (”What if someone finds out about incurable diseases?”)
- users need an explanation of why they get such results. That includes general reasoning as well as the information of what factors influence predictions.
- the presentation of results increased some concerns. Three people voiced concerns that the percentage of likelihood might not be well understood by users. Some suggested having visualisation (e.g. chart) over number. When percentage was changed to descriptive words (e.g. low, high) it was better adopted. Three other people also would prefer to have results as bigger documents with all details.

At the end of the prototyped service, participants were asked for consent to share their data with other organizations. Four participants explicitly appreciated that the default is opt-out. Three participants said that the consent, in general, feels natural now, as many services have it. Still, a few information was missing at this stage:

- (3) reason for sharing results with specific organizations,
- (3) whether the results are shared anonymously,

- (2) short information about the organization,
- (2) how results would be used by other organizations.

Furthermore, three people voiced the need for accessibility: "If I was sitting at my desk and really doing this I would read all these carefully, but with the bigger font of course". Moreover, we discovered that whenever the data or organization to share data with were too briefly described (e.g. stating only a name), they raised emotions of intrusiveness to at least two participants. In general 3 people mentioned that they prefer a bit longer text. Nevertheless, a few of the participants told that they find the level of information good. They would not like to have more text on the page.

A few different attitudes to the service could be differentiated:

1. *I am not interested in most of the text and information presented with some exceptions.* Usually skips some parts and reads only interesting ones, which are most usually for sharing some of the personal data or results. Agrees to most of the things without much thinking. Is used to different services where they share their data and consent to its use.
2. *I am generally interested.* Reads through the whole service. Clicks only on a few specific interesting links to read more. Goes carefully through the data to be shared. Is mostly interested in reasoning and the service process.
3. *I want to know everything.* Opens everything that can be opened. Asks questions before seeing the answer on it. Needs much information to be satisfied with the service.

When talking about trust, I noticed that people are usually referring to two different types of trust: one in the results and another in the service. The former makes users find the results reliable and take suggested actions. For ensuring this trust following actions were mentioned:

- (3) more personal experience with the system, seeing that it works on personal example,
- (2) heard or read statements from other users,
- (2) conversation with a doctor about the results,
- (2) review by a medical professional.

The latter trust, e.g. in the service provided, is strongly related to the trust in the public sector. This trust makes people use the service, even if they will not be able to trust the results or if they miss some information. Having said that, the lack of transparency in services lessens and harms the trust in them. Moreover, one of the youngest participants noticed that "it is difficult to differentiate fake from real those days". The factors which were mentioned for bringing trust during user testing were:

- (6) previous experience and trust to the public sector,
- (4) standard look that resembles other public sector services (e.g. sterile, neutral look that does not shape user feelings)
- (2) organization information.

Regarding the security of the data and service, one person mentioned that it comes from the Public Sector organization and another that it comes from the standard bank login.

4.4 Personas and guidelines: practical outcome of the empirical study

In the following section, I present results of gathered knowledge through empirical study and literature review: personas and guidelines. Personas represent future users of AI PS services. They can be used by designers and developers to understand citizens - users of future AI services. That, in the future, can result in better and more relevant services. Guidelines are containing a set of recommendations that aim to help public sector designers and developers in creating trustworthy AI services.

Personas are based on the data gathered during the empirical part of the study, that is interactions with around 30 Finnish residents. The first action for creating the personas was dividing participants of interviews and user testing into five groups, a few people each, based on age and education. For each of the group, I looked for attitudes, needs, concerns or behaviours that were characteristic only or mostly to that group. The quotations used in personas are also based on what real participants said during interviews. All of those data was later divided into different blocks, described in the table 4.7. Personas can be viewed in appendix D.

It is suggested to use personas during the design stage of creating the new AI service [75]. In the beginning, the bios of personas can help service designers in understanding and relating to future users [74]. Moreover, the needs and concerns parts can help in producing a list of requirements for the product. During the whole design process, personas can keep designers focused on the users' goals and needs, and hence ensure that created service would be user-centric [75]. Finally, personas can serve during the initial testing stage. Before the real users are invited, designers can test their services from the personas viewpoints and find potential flaws [75].

Guidelines are a set of advice for creating trustworthy AI services of the public sector. Firstly, they were created based on the input from participants of the study. Specifically, they addressed citizens' concerns and needs gathered through interviews and design workshop. Moreover, they included citizens' transparency vision from the design workshop. This vision includes ideas for

Content	Description
Hashtags	Two adjectives representing the persona
Information	Data about persona age, education, profession and AI awareness
Bio	An introductory paragraph that aim for better understanding and empathy to persona
Attitude towards the use of AI in ps	Persona's opinion on the potential AI services of the public sector
About Human in the loop	Persona's opinion on what should be the role of human in the AI public sector services
Needs and/or concerns	A short list of needs and/or concerns, typical for the persona
Using the service	A way of interacting with the digital public sector service

Table 4.7: Description of the personas format.

how to present the information needed for trust in a relevant and understandable manner. Those guidelines were next tested and adjusted in the iterative process of user testing. Firstly, the prototype of the public sector AI service was created based on initial guidelines. Then, the prototype was tested with citizens of Finland. Last, the prototype and guidelines were corrected based on the output and tested again. Last but not least, during the whole process of creating guidelines, they were also being aligned with expert guidelines from the literature review.

In the current form, guidelines are divided into two parts: Information Transparency and Principles. The former presents the information needed on different stages of the PS AI service and the priority of the information. It also includes a recommendation on how such information should be presented. The latter section is presenting recommendations to include during service design, development and operation stages. The more detailed structure and content example of the guidelines are presented in table 4.8.

The guidelines can be used in multiple stages of service creation. On the very beginning stage of planning and designing the service, I would recommend considering principles from the section Service Design. Then, when the technical and architectural planning is started, the guidelines from Service development and operations can be used. Next, when the interface of the service is being designed, I would suggest using the whole Information Transparency section as a pattern. Lastly, all guidelines in the document can be used as a checklist after finishing the first service version, whether all points are covered there. The full guidelines document can be viewed in the appendix F.

Section	Description	Example of included guidelines
Information Transparency	Suggests on how to achieve transparency of PS AI services, mainly what information, when and how to present it. Is divided into four stages that represent the citizens' interaction with the service.	*Below example of what type of information should be visible at given stage.*
Information stage	A citizen can get informed there about the service, its purpose and any details. That stage should be accessible at any point.	Detailed process of the service, technical documentation, way of mitigating discrimination and bias, privacy statement
Application Stage	A citizen comes to this stage, when they decide to use the service. That's where the user can fill in any data, preferences or give consent.	General description of the process, that AI will be used, purpose, data used and its sources, security and privacy policy in short
Waiting Stage	In case if the service does not produce immediate results, there is a waiting stage where the user is informed about the current state of the process.	Update on the process, data used
Results Stage	Where a citizen is given results of the service.	General description of the process, explanation of results, data that was used, redress contact
Principles	Groups guidelines of what factors to include or consider while designing, developing and running the service, and motivates for additional activities.	*Below example of principles to be considered in each stage*
Service design	Actions to be considered on the level of planning the service and its operations.	Give citizens control over their data and AI, give citizens an option to interact with a human, make service efficient and beneficial
Service development and operations	Factors needed to be addressed during the technical and architectural development, as well as the operation of the service.	Involve human in the process, develop secure and reliable system, provide privacy, create accessible interface, ensure data quality
Additional activities	Suggests actions that can be taken apart from the service creation processes.	Provide education, keep citizen in the loop

Table 4.8: The description of the content of guidelines.

Chapter 5

Discussion

In the following chapter I discuss the results of this work as well as the thesis in general. The first four sections 5.1, 5.2, 5.3 and 5.4 present the discussion to the research questions one to four accordingly. The last section 5.5 presents the limitations of the work as well as suggestions for the future research.

5.1 Current attitudes towards and concerns about AI use in public sector

In collective people were voicing almost twice more positive than negative attitudes towards AI use in Public Sector. It can be viewed as a good starting point, however, one shouldn't disregard those one in three negative voices. On the contrary, those would need a bigger attention and action taken. Specific attitudes could be divided into five types. Each of those types is also represented in the persona (check appendix D). Those attitudes are:

1. demanding enthusiastic - mostly among young educated people of lower AI awareness,
2. carefree neutral - mostly among young lower educated people,
3. suspicious negative - mostly among more senior lower educated people,
4. cautious calm - mostly among more senior educated people,
5. demanding anxious - only among younger, higher educated and having high awareness of AI group.

The most negative attitudes are seen in groups 3 and 5. The distrust in the AI of the third group might come from the lack of education in modern technology, or from the image created by media. In fact, the only people who admitted to being scared of AI, were of low education and AI awareness levels. The fifth group are people who are young specialist in AI subjects. In their case, the negative attitude might come from the distrust to people who are

providing AI solutions. This group of specialists was aware of already existing cases of unethical AI use, as well as, possible misuses that can happen with it.

The more detailed attitudes were dependent on the person's experience and the context of AI use. The participants were the most positive about the health-related case. The automated decision case, where AI would decide on whether an applicant receives the flat or not, was rated positive by higher educated people. As a reason, they said that AI decisions would be faster and not human-biased. A few people of lower education were negative to this case, feeling that the process is harsh and inhuman. Two cases that received the most negative opinions, were the ones that would need to use data from the private sector (case 4) or about children (case 3). Those cases can indicate what is the privacy boundary for citizens' data use.

One factor joining almost all participants in the empirical study was their assurance of having trust in the public sector. For example, they would allow the public sector to do more with their private data than they would allow the private sector. Some mentioned that it is due to the rules and regulations that the public sector is obligated to obey, its righteous history and no commercial interest. Nevertheless, trust in the public sector does not guarantee trust in the AI. Furthermore, very often every assurance of trust was followed by the list of concerns appearing when talking about AI in the public sector.

Concerns mentioned by the participants were much depending on their education level or age. Still, for most participants of this study, one of the biggest concern was security: whether the data will be safe, whether PS has capabilities to build reliable systems or if technology providers are trustworthy. Citizens were also worried about their privacy, especially when it comes to the data about their relatives and relations. Leaving citizens no privacy could easily lead to becoming the surveillance country, which the participants were the most concerned about. Moreover, they mentioned such as errors of the system or having no control of the AI. Another concern, misuse of the AI or data, was mentioned only by immigrants in Finland. They explained, that his concern comes from their experience with other countries. Examples of such case can be using data to control citizens or monetizing it by sharing with third parties.

Most of the results were alike with the literature study. The collective attitude agrees with the results of the study done on New York times articles [2], where there were twice as many positive opinions about AI than negative ones . Talking specifically about AI used in the public sector, the collective approach would be cautiously positive, same as found in the survey from 2019 [13]. Moreover, similarly to the studies done by RSA[30] and a Nordic company [51], the participants of this study were the most positive about the health-related case.

On the other hand, some inconsistencies exist between results of this study and literature review. The reason for those might be different naming, study environment or used cases. For example, the case that uses children's data for impact assessment of the education system, can be labelled as a research case and it was not accepted by most participants. In the research of Hyry [27],

however, Finnish people are broadly supporting using AI for science. Moreover, concern about bias and discrimination was mentioned by multiple research papers and reports. In the study it was discussed only once among design workshop participants. It was also discovered, that most of the participants are not aware of this problem. Last but not least, results from the literature review adds the concept of underusing of AI [17] to the misuse one.

In summary, citizens keep the cautious positive approach to the idea of public sector AI systems. Results of the empirical study showed that existing trust in the Finnish government can make citizens more likely to use public sector services. However, the concerns of privacy, security and loss of control should be addressed. Moreover, the lower educated and aware of AI citizens are, the more likely they would be careless of even fearful of it. This can make such citizens not willing to use and obey AI services.

Moreover, personas (see appendix D) can be used as a way to relate to citizens and understand needs and perspectives. This is an important stage of the design process [74]. As mentioned in the first paragraph, personas are representing five different attitudes to AI used in public sector, which were distinguished during the study. Moreover, they include four more information pieces. First, they mention the persona's perspective on what should be the role of human and AI in the public sector AI services. Next, they list the persona's most important concerns and needs towards the topic. Third, they explain how that persona would be using the digital AI service provided by the Public Sector. Last but not least, it includes the bio and basic information about the persona, that can help the reader empathize better with them.

5.2 Information about the public sector AI services needed for citizens' trust

Information about the AI service is directly connected to the transparency term: the more information is given, the more transparent the service is. However, not all information has the same importance for citizens. During the study, participants were addressing transparency a lot, however, each of them had a bit different perspective on what it is and what it should contain. In general, they were mentioning that they need an easy access to all information about the service that is relevant for them. The participants of the study mentioned that transparency is important to make more informed decisions or to be less suspicious about the service. Moreover, they saw that transparency can lessen customer service workload in the public sector. The more citizens will know from service itself, the fewer questions they will be asking.

The need for an explanation of AI system actions or decisions was the most often repeated one in the following study. The participants of this study pointed out that they need to know why specific results were provided, because it can help them in understanding, accepting or redressing them. Moreover, they needed to know what criteria was used for a specific decision. Sec-

ond most needed information was about the data used in the service. Participants were mentioning their need to know of which of their data is used, how is it collected and from which source.

Next, often mentioned information group, is, as one of the participants called, 3-p: process, purpose and performance. About the process, the participants wanted to learn in a concise way, how the data is used in the service and what is the role of AI and human in it. Regarding the purpose, they wanted to understand why should they use provided service, why was it created and how it can benefit them or other people. Last of the three, performance, was mentioned mostly by people of bigger AI knowledge. They mentioned, that they find it important to inform how accurate are the results. According to the participants, those pieces of information can in deciding whether they want to use the service and how much should they trust it.

Pieces of information, that were less often mentioned by participants, were following. First, a few participants pointed out that in a service it should be clear what public institutions are providing the service. It also should be clear how to contact it in case of redress. Next, the participants voiced their interest in understanding the impact of the service. That was especially appearing in cases involving relatives or automatic decision making. Last but not least, some participants also pointed out that they would need recommendations on what they should do with given results.

The importance of transparency is also visible in many studies used in the literature review of this thesis [4, 10, 31, 55]. Even more, those studies are mainly focusing on the same pieces of information that study participants mentioned. Such repeated information is for example process, purpose, performance, used data or the impact of the service.

The need for an explanation was also emphasized in the literature review. It seems that it became an important topic in the research branch of AI. However, it is important to mention, that explanations in AI have a double meaning in those research materials. First, technical is more focused on algorithmic explanations of the AI actions. Second, societal meaning focuses on the justification of AI actions in an understandable for general users way [56, 58, 66]. It is the latter that was chosen by both participants of the study and some experts as the most important, when talking about trustworthy services [28, 57, 59].

The need for an open code or algorithms used in the service was a type of information that was much more often mentioned in the literature review [6, 11, 63, 64], than in the empirical study. It is important to add, that before the code is published, it should be reviewed, whether sharing it would not create risks of, for example, hacking the system [37, 62]. In the study, only a few participants who are having good AI knowledge shared their interest in getting to know the used algorithm, mostly out of curiosity motives.

5.3 Factors that are needed for building citizens' trust towards Public Sector AI services

The needs voiced by citizens in the following study regarding public sector AI services were alike the ones in the literature review. The service factors that gathered all together the most interest are being in control over AI and data; and having a human involved in the whole process of the service.

The former, the user control, was reflected in many different ways during the study. One of examples were, when participants shared their feelings of helplessness towards what is being done with their data. The citizens from this study were requesting to always be asked for the consent before their data is being used. They would want to be able to have full control of it and be always able to access it, which was also supported by multiple existing papers and guidelines [38, 56, 62]. They mentioned, that the access to data could also help them to notice the situation when data stored is outdated. Moreover, participants wouldn't like data to be shared with other organizations without their knowledge and acceptance. However, it is known that in some cases the public sector might already have the right to use specific citizens data. In such a situation, citizens should be still informed beforehand about the planned data use and the law in use. Apart from data control, participants also voiced the need to be in control of AI. They suggested, for example, that they want to be able to not use AI or even ask for human work instead of AI, same as advised in European guidelines [28].

The latter, about human involvement, can be divided into two sections that are described in European guidelines [28] as human-on-the-loop and human-in-command. The first approach explains the need of having human overlooking the AI system behaviour. Possibly audits or certifications should be enabled for informing about it. Next, human-in-command means that human should be responsible for decisions, either by using the AI system as an advisor or by reviewing the decision before handing it to the end-user. Many participants were mentioning that the knowledge that an expert reviewed the decision can help them trust it, especially in healthcare cases, which is also confirmed in the studies done in Finland [27] and in Nordics [51]. On the other hand, it was mentioned by a few participants of this study that decision making by AI can be better than human-made due to lack of corruption, human bias and human error, which was also confirmed by the study done by RSA [30]. Lastly, human-in-command is about having interactive tasks handled by a human [7]. The participants mentioned that tasks such as negotiating or handing in the personal results, for example, healthcare-related are better to be done with person, which can help in avoiding misunderstandings and the feeling of being dehumanized.

The need for privacy, security and reliability of AI systems used in the public sector services, was mentioned more as a concern during the empirical study, while in literature review it was one of the most repetitive need

[4, 37, 52]. Nevertheless, it all shows the importance of citizens being assured of their safety. Any data leaking, unethical data sharing or unreliability of services, especially those handling citizens personal information, can cause a dramatic drop in trust towards the public sector. Therefore, it is important for the public sector to proactively manage risks, be prepared for any failures and openly speak about those [4, 28]. That also includes the risk of bias and discrimination, broadly described in the review literature review [4, 6, 55, 64]. The possible reason for those topics being hardly mentioned during the empirical study, can be lack of education and being a part of less-likely to be discriminated groups. The only people who mentioned this problem were ones, who are actively working with AI. Lastly, as mentioned in the multiple guidelines [4, 11, 38, 57, 62], it is important that in case something happens, there is an accountable person selected.

Participants in the study pointed out also a need to comprehend the benefit of the service. That was especially noticeable when comparing results between two cases presented in the interview. The first was providing helpful health prediction, while the second was mentioning about the education impact assessment project based on tracked data from school pupils. The benefit of the first case was much more understandable and tangible, therefore the case gained much more positive opinions. Some participants explicitly mentioned, that they need to know what is the benefit for them before they will want to use it. A few guidelines [8, 17, 28, 62] add the need for providing the benefit to the society or environment, as well as always respecting citizens right and democratic values. Those were not explicitly mentioned in the study, possibly thanks to the up-to-day positive experience with the public sector in Finland.

Another topic was broadly mentioned in the literature review and less often in the empirical study, and yet it is an important factor specifically for public sector services: services need to be accessible to all citizens [33, 63, 68]. In the case of AI-using services, the interfaces have to be clear, easy and predictable [56]. The language must be understandable by different citizens groups and service possible to accessed from every places and community [62, 69]. Moreover, during the design workshop, participants were mentioning about the importance that the public sector interfaces are clean and not manipulating anyhow its users. Moreover, they suggested, that public sector interfaces could be similar between each other in appearance, to bring the feeling of familiarity and therefore trust.

The last topic is about providing good education and including citizens in the dialogue about AI. It is not a factor needed in the design of the services, but rather something that the public sector could be responsible for as a side action. The need for education can be seen from studying attitudes. There, feelings like resistance, fear or helplessness appear out of the lack of knowledge. Moreover, several participants were mentioning that they would gladly learn more about the topic of AI. It is important to mention, that probably this does not mean that everyone should receive a technical education in the topic of AI. Rather, citizens could be taught about the basics of what AI is and what

benefits and risks it can bring [6, 7, 17, 38]. This could help in clearing false stereotypes. On the other hand, it is also important to provide social and ethical studies for technical students [17]. That could help in future in avoiding creating unethical AI products.

Finally, involving citizens in decisions related to AI use in the public sector is recommended as an active AI awareness-building [4, 11, 13, 28]. Two of the study participants were mentioning eagerness to be included in shaping Finnish AI policy. That can be done by, for example, public debates. It can not only satisfy the needs of citizens but also enable the public sector to do more informed and publicly acceptable decisions.

Last but not least, it is important to mention the factors that are not under the control of the public sector. Those are word of mouth, that is what friends, family or media are saying [46, 67], and previous experience with AI or with the public sector organization [23]. For example, existing good experience with the public sector in Finland enabled user testing participants to use the prototyped service, even when they had some concerns about it. The trust to use the services can also come from the knowledge that other citizens are also using the service. On the other hand, the participants mentioned that the trust to the results from AI is more likely to increase when they see by themselves or hear about cases where those results were correct. That can also be used as an argument for public sector transparency for using AI: the more citizens will hear about positive usage about AI, the more likely they will want to use it themselves.

5.4 Guidelines for design and development of trustworthy AI services

Guidelines for the design and development of trustworthy AI services might be considered as the main result of this work from the perspective of practice. They are context-based, as advised by Rostlinger and Croholm [33], where the context is of public sector organizations that would like to introduce AI to their services. Moreover, those guidelines are based on thorough literature research and discussion with people of different backgrounds, as recommended by Thomson et al. [71]. Finally, they fill in the gap of lack of empirical studies and citizen view on the AI usage in the public state [17, 30, 34]. In fact, the big part of guidelines structure and content comes from multiple interactions with citizens.

There were no clear instructions found on how to structure guidelines. Therefore, its structure was decided by the iterative process of three steps. First, gathering all needs, relevant principles, guidelines and concerns. Second, clustering them with similar topics. Third, testing and improving, based on results from the user testing. In the final step, I decided to divide the guidelines into two main sections: Information Transparency and Principles. The sections are described below and the full guidelines document can be viewed

in appendix F.

The first part of the guidelines is about Information Transparency. It was structured with an aim to help designers to provide understandable and usable information, the need pointed out by Abdul et al. [34]. The section is divided into four parts, which represent different service stages: informative, application, waiting and results. In each, it is suggested what information should be visible there and of how big priority it is. The bigger priority, the more visible the information should be. The lower - it should not grab immediate attention, but they should still be accessible.

The informative stage of the service could be usually the first place that citizens would interact with in the service. It can, for example, be a publicly accessible web page. We suggested in our prototype, that such a web page would contain a set of links to articles. The link titles would be presented as questions, some of them both more general, introducing to the service and other more detailed. Such links to articles could be grouped by topics to help citizens finding the most relevant information quickly. Examples of topics that should be included there are: process explained in better detail, therein exact role of human; technical documentation, therein performance of AI system, data quality, the security of the system and data storage; whether the system is monitored, certified or audited; (If possible) open-sourced code or Impact of the service on user and society, e.g. visualisations of demographic. The informative stage of the service should be accessible from any other stages.

The next stage, application, represents the moment in the service where a citizen is decided to start using the service. Here, it is important to repeat the most important information about the service: process, whether AI will be used and its purpose. Comparing to the information stage, here that information should be presented in a shorter and more concise way, in order to increase the chances of a user reading those. Moreover, in this stage, it was important for the participants to understand what of their data will be used, what are its sources and what are the privacy and security policies in short. If possible, citizens would prefer to give consent on sharing specific data. Importantly, not every citizen is interested in knowing all details about data, therefore in our prototype we suggested drop-down lists, that would allow citizens to choose how much they want to see.

The third stage is when citizens would need to wait for the results of their participation in the service. Therefore, this stage would be not relevant in cases such results would come immediately. In this stage, citizens wanted to be informed about the current update on producing the results. They would also want to be able to modify their data sharing consents or even opt-out from the service.

The last stage we included in guidelines is the results stage. There, results should be presented in an understandable way, such as decision information, impact assessment results or predictions. It is most important to present the explanation for why such results were created. Moreover, short information reminder about the process and data that were used for this decision should be

also highlighted. Several more pieces of information, such as a way of redress, data safety or recommendations are also suggested to be included there.

The need for creating interfaces that are not overloaded, nor underloaded with information came from interactions with citizens, mainly during user testing. Most of them wouldn't like to read the plain long text placed upon the service page. It could make them feel overwhelmed or disinterested, which could lead to not using or misusing the service. On the other hand, when the information was too scarce, it was creating the feeling of intrusiveness. Moreover, the placement of when different information should be accessible came from mainly design workshops and was confirmed or updated by user testing.

The next part of the guidelines is called principles. It is divided into two sections which are representing different stages of producing the new service: service design, development and operating. In each section, there are guidelines presented that should be considered during the specific stage. As suggested by Amershi et al. [56], the guidelines are presented as short and actionable instructions.

The service design stage happens when the idea for service appeared and is turned into a more planned concept. Some of the most important recommendations for this time would be about making sure to include some factors in the service design. Such factors could be giving citizens control over their own data and over AI actions, or making sure that the service would be beneficial and respecting citizens rights.

The next stage tackles the period when the concept is being developed into working service and sustained. There, it is important to ensure that areas such as security, reliability or privacy of the service are meeting requirements. Moreover, it is suggested to include human in the service processes, either as auditor or part of the process. Other guidelines tackle such topics as providing a good interface, mitigating bias and ensuring data quality. As additional activities, recommendations are added for supporting education and keeping citizens in the loop.

Apart from the guidelines themselves, this work provides two other practical materials that can help service creators: personas and the service prototype. The former can be used as a way to understand the users and make sure that the service would satisfy them. The latter can serve as a tested example of how to place different information on different service stages. Following those materials together with guidelines, would help service creators to answer citizens concerns and needs, as well as ensure that their service meets ethical standards. Henceforth, it would ensure the trustworthiness of those services.

5.5 Limitations and future research

The first important limitation of this study is that it was done with a small representative of Finnish residents. For a full understanding of the existing attitude, concerns and needs in Finnish society, more groups should be included,

for example, teenagers or elderly. Moreover, IEEE guidelines [38] suggest to identify and include cultural minorities in served society, as they would also be served with those new services.

Furthermore, I would not advice to directly apply the results of this study in other cultures. More research would need to be done, as in this thesis it is partially focused on the Finnish environment. Moreover, the findings would need to be verified [19], by, for example, user testing the service prototype. There are multiple reasons for those actions. In different parts of the globe, societies have different cultural biases and values, that can influence the mental model of AI [23]. Furthermore, countries have different focuses on different technologies and societies are differing in the use of different tools, which can influence the adoption of the new AI services [49]. Even within Europe, the cultures and social characteristics are different between each other [19, 35]. Last but not least, Finland is different comparing to other countries thanks to its greater trust in the public sector [12]. Even during the interviews, a few participants that had experiences with other countries mentioned that in Finland they would accept more than in other places.

In this work, the main focus was on interacting with citizens. Therefore, there were no resources left to test the guidelines with designers and developers. Such tests, however, are much needed to understand whether those guidelines are understandable. Some case studies could be done, where presented guidelines would guide the full design and development of a new public sector service. This would help in evaluating whether guidelines can indeed provide help and good results.

Another limitation of this study is that there is no one clear definition of AI that would be common between studies [4, 77]. Moreover, there are yet little studies on societal needs towards AI applications in the public sector. Therefore, materials that were used in the literature review were focusing on slightly different technologies: automation systems, automated decision systems, machine learning, artificial intelligence. That could have brought results that might not be relevant to all AI applications in the public sector.

Furthermore, the AI applications itself could have been more specified. Only one case could be chosend out of, for example, automated decision, citizen assistance or impact assessment. As this study showed, people are reacting differently to different AI-usage cases. Moreover, the context of the application, how it can influence one life (for example whether it is used in the healthcare or in the post services) also influences the level of information and factors that should be applied there [20, 37]. Having focused on one type of the AI application the final guidelines could be even more usable and relevant.

Moreover, via different studies, as well as empirical research it was discovered that there are two different instances of trust relevant when talking about trustworthy services: trust to the service provider and trust to the AI. The former can help citizens in making the decision to use provided services, for example, to share the data with it. The latter makes them accept the results and follow the given recommendations. In this study, there was no focus on

one of those trust instances. It could be valuable to research on the importance and current state of each, as well as what factors are influencing which of those instances.

The main goal of this work was to provide guidelines that would build trust. However, the following question remains: how much of trust to the public sector or AI is actually healthy? From one side, the trust can help citizens accept and use tools that majority of us do not fully understand, like 4G network [22]. On the other hand, it was seen in this research, that citizens have much less knowledge about data being used in the public sector than what they knew about the private sector, possibly thanks to the higher trust. Having too much of trust can limit citizens' ability for critical evaluation and learning [46, 77]. Moreover, it builds our reliance on AI, which we might not even be aware of [8]. Altogether, it could be worth to explore the proper balance of trust and distrust.

Last but not least, this work focuses more on the factors that in the short term would help in gaining trust. However, more research would be needed to check the longer-term effects of introducing different AI applications in the public sector. The examples of sectors with questions that such research should include are: social - would AI application help in decreasing the social injustice between communities?; economical - who is going to earn money from public sector AI services, especially if such is provided by the third party?; global - how are we affecting other countries by using such a tool?; and environmental - is the value that AI brings bigger than the environmental trace (energy, minerals) that is left?

Chapter 6

Conclusions

This thesis aimed to answer the question of "what needs to be taken into consideration while designing trustworthy public sector AI services?". As a method, the design process was used, that consisted of three stages. In the first stage, over 20 Finnish residents were interviewed about their attitudes, concerns and needs regarding AI use in the public sector. During the next stage, design workshop, Finnish residents brainstormed together about trustworthy public sector AI services. Last, during the user testing, the prototyped public AI service was tested with Finnish residents. Altogether, over 30 residents of Finland participated in this qualitative process. Moreover, the results of this study were enriched with perspectives from the broad literature review. That included the newest scientific publications, technical reports, surveys and existing guidelines. This thesis finishes with three main conclusions, written below.

There are more positive than negative voices about the usage of AI in the public sector, however, the number of the latter is significant. The most negative voices came from senior people of low education and junior AI specialists. The trust in the public sector is strong in Finland, which made participants more eager to use prototyped AI public sector service. Nevertheless, partipants voiced multiple concerns, such as whether the public sector would be able to provide secure services or whether the use of AI in the public sector wouldn't lead to the lack of privacy and surveillance. Based on the interactions with citizens, personas were created (appendix D). Those can help service designers in understanding citizens and relating to their needs.

It is important to keep the public sector services transparent, in order to keep trust to the public sector and build the trust to AI. Citizens need to know when AI is used, how and for what purpose. This way they can build their understanding and experience of AI usage in the public sector. Furthermore, citizens need to know why AI services resulted in specific actions or decisions, as well as, which of their data was used in the process. Such information needs to be presented to them in an understandable and yet concise way. Interfaces cannot be overloaded with information, however, more details should be available for those interested. Specific instructions on how to

present different information about the service are grouped in the first part of Guidelines (appendix F). The example of how such a service could look like is presented in the service prototype (appendix E).

Citizens' needs and concerns, as well as ethical requirements, ought to be addressed in the design and development of trustworthy public sector AI services. The specific recommendations are listed in the second part of the attached guidelines (appendix F). Those are, for example, mitigating security and discrimination risks, providing citizens with control over their data and having a person involved in AI processes. In the end, citizens will be more likely to build trust to the public sector, when they see that their needs and concerns are addressed, through the transparent service, accessible information and the positive experience.

Bibliography

[1] Science and Technology and Committee, "Robotics and artificial intelligence," House of Commons, Tech. Rep., 2016. [Online]. Available: www.parliament.uk/science

[2] E. Fast and E. Horvitz, "Long-Term Trends in the Public Perception of Artificial Intelligence," in *AAAI Conference on Artificial Intelligence*, vol. 31, 2017. [Online]. Available: www.aaai.org

[3] C. Rzepka and B. Berger, "User Interaction with AI-enabled Systems: A Systematic Review of IS Research," *nternational Conference on Information Systems*, vol. 39, 2018.

[4] B. W. Wirtz, J. C. Weyerer, and C. Geyer, "Artificial Intelligence and the Public Sector - Applications and Challenges," *International Journal of Public Administration*, vol. 42, no. 7, pp. 596–615, 5 2019.

[5] *Finland's Age of Artificial Intelligence: Turning Finland into a leading country in the application of artificial intelligence. Objective and recommendations for measures*.

[6] "AI Now Report 2018," AI Now, Tech. Rep., 2018. [Online]. Available: www.ainowinstitute.org

[7] High-Level Expert Group on AI, "Policy and investment recommendations for trustworthy Artificial Intelligence — Shaping Europeâs digital future," European Commission, Tech. Rep., 4 2019. [Online]. Available: https://ec.europa.eu/digital-single-market/en/news/policy-and-investment-recommendations-trustworthy-artificial-intelligence

[8] C. Cath, S. Wachter, B. Mittelstadt, M. Taddeo, and L. Floridi, "Artificial Intelligence and the âGood Societyâ: the US, EU, and UK approach," *Science and Engineering Ethics*, vol. 24, no. 2, pp. 505–528, 4 2018.

[9] T. Q. Sun and R. Medaglia, "Mapping the challenges of Artificial Intelligence in the public sector: Evidence from public healthcare," *Government Information Quarterly*, vol. 36, no. 2, pp. 368–383, 4 2019.

[10] H. Mehr, "Artificial Intelligence for Citizen Services and Government," Harvard Ash Center Technology & Democracy, Tech. Rep., 2017.

[11] M. L. Smith, M. E. Noorman, and A. K. Martin, "Automating the public sector and organizing accountabilities," *Communications of the Association for Information Systems*, vol. 26, no. 1, pp. 1–16, 2010.

[12] Ministry of Economic Affairs and and Employment, "Leading the way into the age of artificial intelligence Final report of Finland's Artificial Intelligence Programme 2019," Helsinki, Tech. Rep., 6 2019.

[13] M. Carrasco, S. Mills, A. Whybrew, and A. Jura, "The Citizen's Perspective on the Use of AI in Government," BCG Digital Government Benchmarking, Tech. Rep., 3 2019.

[14] AI Now Institute, "AUTOMATED DECISION SYSTEMS Examples of Government Use Cases," Tech. Rep., 2019.

[15] Colin Lecher, "New York Cityâs algorithm task force is fracturing," 2019. [Online]. Available: https://www.theverge.com/2019/4/15/18309437/new-york-city-accountability-task-force-law-algorithm-transparency-automation

[16] G. Adamson, J. C. Havens, and R. Chatila, "Designing a Value-Driven Future for Ethical Autonomous and Intelligent Systems," *Proceedings of the IEEE*, vol. 107, no. 3, pp. 518–525, 3 2019.

[17] L. Floridi, J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Luetge, R. Madelin, U. Pagallo, F. Rossi, B. Schafer, P. Valcke, and E. Vayena, "AI4People - An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations," *Minds and Machines*, vol. 28, no. 4, pp. 689–707, 12 2018.

[18] J. Cowls and L. Floridi, "Prolegomena to a White Paper on an Ethical Framework for a Good AI Society," *SSRN Electronic Journal*, 7 2018.

[19] J. D. Lee and K. A. See, "Trust in Automation: Designing for Appropriate Reliance," *Human Factors*, vol. 46, no. 1, p. 50â80, 2004.

[20] M. Arnold, R. K. E. Bellamy, M. Hind, S. Houde, S. Mehta, A. Mojsilovic, R. Nair, K. N. Ramamurthy, D. Reimer, A. Olteanu, D. Piorkowski, J. Tsay, and K. R. Varshney, "FactSheets: Increasing Trust in AI Services through Supplier's Declarations of Conformity," *Ibm Journal of Research and Development*, 8 2018. [Online]. Available: http://arxiv.org/abs/1808.07261

[21] "Trust and AI," Taiste, Tech. Rep.

[22] M. L. Smith, "Building institutional trust through e-government trustworthiness cues," *Information Technology and People*, vol. 23, no. 3, pp. 222–246, 2010.

[23] K. E. Schaefer, J. Y. Chen, J. L. Szalma, and P. A. Hancock, "A Meta-Analysis of Factors Influencing the Development of Trust in Automation: Implications for Understanding Autonomy in Future Systems," *Human Factors*, vol. 58, no. 3, pp. 377–400, 2016.

[24] "Human-AI Collaboration Trust Literature Review - Key Insights and Bibliography - The Partnership on AI." [Online]. Available: https://www.partnershiponai.org/human-ai-collaboration-trust-literature-review-key-insights-and-bibliography/

[25] N. Antunes, L. Balby, F. Figueiredo, N. Lourenco, W. Meira, and W. Santos, "Fairness and Transparency of Machine Learning for Trustworthy Cloud Services," in *Proceedings - 48th Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops, DSN-W 2018*. Institute of Electrical and Electronics Engineers Inc., 7 2018, pp. 188–193.

[26] M. S. Blumenthal, "The Politics and Policies of Enhancing Trustworthiness," *Communication Law and Policy*, vol. 4, no. 4, pp. 513–555, 1999.

[27] J. Hyry, "The use of digital services," Tech. Rep., 2019.

[28] High-level expert group on Artificial Intelligence, "Ethics Guidelines for Trustworthy AI," European Commission, Tech. Rep., 2019. [Online]. Available: https://ec.europa.eu/digital-

[29] R. C. Nyhan, "Changing the paradigm: Trust and its role in public sector organizations," *American Review of Public Administration*, vol. 30, no. 1, pp. 87–109, 2000.

[30] "Democratising decisions about technology: a toolkit - RSA." [Online]. Available: https://www.thersa.org/discover/publications-and-articles/reports/democratising-decisions-technology-toolkit

[31] M. Grimsley and A. Meehan, "e-Government information systems: Evaluation-led design for public value and client trust," *European Journal of Information Systems*, vol. 16, no. 2, pp. 134–148, 4 2007.

[32] A. Salminen and R. Ikola-Norrbacka, "Trust, good governance and unethical actions in Finnish public administration," *International Journal of Public Sector Management*, vol. 23, no. 7, pp. 647–668, 10 2010.

[33] A. Rostlinger and S. Croholm, "Design criteria for public e-services," vol. 28, 2009. [Online]. Available: http://aisel.aisnet.org/ecis2008/28

[34] A. Abdul, J. Vermeulen, D. Wang, B. Y. Lim, and M. Kankanhalli, "Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda," in *Conference on Human Factors in Computing Systems - Proceedings*, vol. 2018-April. Association for Computing Machinery, 4 2018.

[35] M. M. Rantanen and J. Koskinen, "ETHICAL FRAMEWORK FOR A FAIR, HUMAN-CENTRIC DATA ECONOMY WP 1: Citizens' values report for IHAN," Tech. Rep., 2019.

[36] "A Consortium of Finnish organisations seeks for a shared way to proactively inform citizens on AI use," 6 2019. [Online]. Available: https://www.espoo.fi/en-US/A_Consortium_of_Finnish_organisations_se(167195)

[37] Treasury Board of Canada Secretariat, "Directive on Automated Decision-Making," Tech. Rep., 2 2019.

[38] *ETHICALLY ALIGNED DESIGN A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*, First edition ed. The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systemsa, 2019.

[39] C.-H. Tsai and P. Brusilovsky, "Designing Explanation Interfaces for Transparency and Beyond," *IUI Workshopsâ19*, 2020.

[40] Design Council, "What is the framework for innovation? Design Council's evolved Double Diamond — Design Council." [Online]. Available: https://www.designcouncil.org.uk/news-opinion/what-framework-innovation-design-councils-evolved-double-diamond

[41] S. J. Clune and S. Lockrey, "Developing environmental sustainability strategies, the Double Diamond method of LCA and design thinking: A case study from aged care," *Journal of Cleaner Production*, vol. 85, pp. 67–82, 12 2014.

[42] H. R. Bernard, *Research Method in Anthropology: Qualitative and Quantitative Approaches*, 4th ed. Lanham: AltaMira Press, 2006.

[43] B. DiCicco-Bloom and B. F. Crabtree, "The qualitative research interview," pp. 314–321, 4 2006.

[44] J. Lazar, J. Feng, and H. Hochheiser, *Research Methods in Human-Computer Interaction - 2nd Edition*, 2nd ed. Morgan Kaufmann, 4 2017.

[45] J. Michanek and A. Breiler, *The Idea Agent: The Handbook on Creative Processes*, 2nd ed. Routledge, 7 2013. [Online]. Available: https://www.goodreads.com/book/show/19222020-the-idea-agent

[46] V. Alexander, C. Blinder, and P. J. Zak, "Why trust an algorithm? Performance, cognition, and neurophysiology," *Computers in Human Behavior*, vol. 89, pp. 279–288, 12 2018.

[47] B. J. Dietvorst, J. P. Simmons, and C. Massey, "Algorithm aversion: People erroneously avoid algorithms after seeing them err," *Journal of Experimental Psychology: General*, vol. 144, no. 1, pp. 114–126, 2015.

[48] M. K. Lee, "Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management," *Big Data and Society*, vol. 5, no. 1, 2018.

[49] H. Nitto, D. Taniyama, and H. Inagaki, "Current Status of Social Acceptance of Robots and Artificial Intelligence II Differences in Attitudes toward and Acceptance of Robots in Japan, the U.S. and Germany," Tech. Rep., 2017.

[50] A. C. Elkins, N. E. Dunbar, B. Adame, and J. F. Nunamaker, "Are users threatened by credibility assessment systems?" *Journal of Management Information Systems*, vol. 29, no. 4, pp. 249–262, 4 2013.

[51] "Nordic AI survey AI through the eyes of the consumers," Tieto, Tech. Rep., 2019.

[52] D. Leslie, "Understanding artificial intelligence ethics and safety A guide for the responsible design and implementation of AI systems in the public sector," The Alan Turing Institute, Tech. Rep., 2019. [Online]. Available: https://doi.org/10.5281/zenodo.3240529

[53] M. Turilli and L. Floridi, "The ethics of information transparency," *Ethics and Information Technology*, vol. 11, no. 2, pp. 105–112, 2009.

[54] Organisation for Economic Co-operation and Development, "OECD Principles on Artificial Intelligence." [Online]. Available: http://www.oecd.org/going-digital/ai/principles/

[55] "DISCRIMINATING SYSTEMS Gender, Race, and Power in AI," Tech. Rep. [Online]. Available: http://cdn.aiindex.org/2018/AI%20Index%202018%20Annual%20Report.pdf.

[56] S. Amershi, K. Inkpen, J. Teevan, R. Kikin-Gil, E. Horvitz, D. Weld, M. Vorvoreanu, A. Fourney, B. Nushi, P. Collisson, J. Suh, S. Iqbal, and P. N. Bennett, "Guidelines for Human-AI Interaction," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*. New York, New York, USA: ACM Press, 2019, pp. 1–13. [Online]. Available: http://dl.acm.org/citation.cfm?doid=3290605.3300233

[57] "AI Principles - Future of Life Institute." [Online]. Available: https://futureoflife.org/ai-principles/

[58] "General Data Protection Regulation (GDPR) - Official Legal Text." [Online]. Available: https://gdpr-info.eu/

[59] P. Pu and L. Chen, "Trust-inspiring explanation interfaces for recommender systems," *Knowledge-Based Systems*, vol. 20, no. 6, 2007. [Online]. Available: www.elsevier.com/locate/knosys

[60] "A guide to using artificial intelligence in the public sector - GOV.UK." [Online]. Available: https://www.gov.uk/government/publications/understanding-artificial-intelligence/a-guide-to-using-artificial-intelligence-in-the-public-sector

[61] "Personal information charter - Department for Work and Pensions - GOV.UK." [Online]. Available: https://www.gov.uk/government/organisations/department-for-work-pensions/about/personal-information-charter

[62] "The Declaration - Montreal Responsible AI." [Online]. Available: https://www.montrealdeclaration-responsibleai.com/the-declaration

[63] "Service Standard - Service Manual - GOV.UK." [Online]. Available: https://www.gov.uk/service-manual/service-standard

[64] "Discrimination, artificial intelligence, and algorithmic decision-making," Tech. Rep.

[65] "Government Design Principles - GOV.UK." [Online]. Available: https://www.gov.uk/guidance/government-design-principles

[66] "Principles for Accountable Algorithms and a Social Impact Statement for Algorithms :: FAT ML." [Online]. Available: https://www.fatml.org/resources/principles-for-accountable-algorithms

[67] R. Heintzman and B. Marson, "People, service and trust: is there a public sector service value chain?" *International Review of Administrative Sciences*, vol. 71, no. 4, pp. 549–575, 12 2005. [Online]. Available: http://journals.sagepub.com/doi/10.1177/0020852305059599

[68] "Finland Principles of Digitalisation." [Online]. Available: https://oecd-opsi.org/toolkits/finland-principles-of-digitalisation/

[69] B. Shneiderman, C. Plaisant, M. Cohen, S. Jacobs, N. Elmqvist, and N. Diakopoulos, "Grand challenges for HCI researchers," *Interactions*, vol. 23, no. 5, pp. 24–25, 9 2016.

[70] S. Cronholm, "The Usability of Usability Guidelines-a Proposal for Meta-Guidelines," Tech. Rep. [Online]. Available: http://portal.acm.org/dl.cfm

[71] R. Thomson, M. Lavender, and R. Madhok, "Fortnightly Review: How to ensure that guidelines are effective," *BMJ*, vol. 311, no. 6999, p. 237, 7 1995.

[72] Villani Cedric, "FOR A MEANINGFUL ARTIFICIAL INTELLIGENCE Mathematician and Member of the French Parliament," Tech. Rep., 2017.

[73] "What are Design Guidelines? — Interaction Design Foundation." [Online]. Available: https://www.interaction-design.org/literature/topics/design-guidelines

[74] J. Pruitt and J. Grudin, "Personas: Practice and Theory," Tech. Rep., 2003.

[75] F. Y. Guo, S. Shamdasani, and B. Randall, "Creating effective personas for product design: Insights from a case study," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 6775 LNCS. Springer, Berlin, Heidelberg, 2011, pp. 37–46.

[76] G. Randolph, "Use-Cases and Personas: A Case Study in Light-Weight User Interaction Design for Small Development Projects," *Informing Science: International Journal of an Emerging Transdiscipline*, vol. 7, pp. 105–116, 2004.

[77] "Human-AI Collaboration: Trust Literature Review - Key Insights and Bibliography - The Partnership on AI," Collaborations Between People and AI Systems, Tech. Rep., 9 2019. [Online]. Available: https://www.partnershiponai.org/human-ai-collaboration-trust-literature-review-key-insights-and-bibliography/

Appendix A

Interview questions

Interview #____ Date & time _____

Demographics

Gender:

Age:

Level and sector of studies/education - Koulutustausta:

Current Occupation - Nykyinen ammatti (asema):

Nationality + how long in Finland if not Finnish:

General knowledge - Yleistiedot

What do you know about the use of your personal information in the private sector? Minkälainen käsitys sinulla on omien tietojesi käytöstä **yksityisen sektorin** palveluissa?

(if the person confused, eg. showing recommendations in the online shop based on your shopping history)

How do you feel about such data usage? Why? Mitä ajattelet tietojen käytöstä? Miksi

What sort of things are ok and what are not ok? Why? Minkälaiset asiat ovat mielestäsi ok ja mitkä eivät ole ok?

What do you think about the use of your personal data in the public sector? Minkälainen käsitys sinulla on omien tietojesi käytöstä **julkisen sektorin** palveluissa?

(if the person is confused, eg. sharing your basic data based on your social security number)

How much do you know about it? Minkä verran luulet tietäväsi?

Where and how could you find out how it's used? Mistä ja miten saisit selville, miten tietojasi käytetään?

How would you describe Artificial Intelligence? Miten kuvailisit tekoälyä - mitä se on?

What would be your definition? Mikä olisi sinun määritelmäsi?

What sort of things are AI and what are just basic IT? Minkälaiset asiat ovat tekoälyä ja mitkä perus ATK:ta

What is the border line between those?

The user stories and use cases we present to you are fictional. They represent a possible future but they do not exist at the moment and some of them may never be implemented.
Seuraavat tilanteet ja käyttäjätarinat ovat täysin fiktiivisiä. Ne kuvaavat mahdollista tulevaisuutta, mutta niitä ei ole nyt olemassa eikä niitä välttämättä koskaan tällaisina toteuteta.

Case #_____

What are your feelings about the use case? Why? Mitä ajattelet tästä tarinasta ja tilanteesta? Miksi?

What questions does it raise? Minkälaisia kysymyksiä se herättää?

What do you think: what kind of data was used there? Minkälaista dataa tässä on käytetty?

Where did it come from? Mistä ja miten se on saatu käyttöön?

What would you like to know about processes used for this action? Mitä tietoja kaipaisit tällaisessa tapauksessa? Missä ja miten tuo tieto pitäisi olla saatavilla?

Eg. How was the decision made, level of AI autonomy, based on what data…

Case #_____

What are your feelings about the use case? Why? Mitä ajattelet tästä tarinasta ja tilanteesta? Miksi?

What questions does it raise? Minkälaisia kysymyksiä se herättää?

What do you think: what kind of data was used there? Minkälaista dataa tässä on käytetty?

Where did it come from? Mistä ja miten se on saatu käyttöön?

What would you like to know about processes used for this action? Mitä tietoja kaipaisit tällaisessa tapauksessa? Missä ja miten tuo tieto pitäisi olla saatavilla?

Eg. How was the decision made, level of AI autonomy, based on what data…

Case #_____

What are your feelings about the use case? Why? Mitä ajattelet tästä tarinasta ja tilanteesta? Miksi?

What questions does it raise? Minkälaisia kysymyksiä se herättää?

What do you think: what kind of data was used there? Minkälaista dataa tässä on käytetty?

Where did it come from? Mistä ja miten se on saatu käyttöön?

What would you like to know about processes used for this action? Mitä tietoja kaipaisit tällaisessa tapauksessa? Missä ja miten tuo tieto pitäisi olla saatavilla?

Follow - up questions - Loppukysymykset

Kun ajatellaan näitä esimerkkejä, niin millä tavalla haluaisit, että niistä kerrotaan?

Minkälainen rooli sinulla pitäisi olla näissä tapauksissa, mitä asioita pitäisi pystyä seuraamaan tai kontrolloimaan? Missä vaiheessa? Miksi? (Suostumus, tarkistus, hyväksyntä yms.)

Mitä näissä käytetyistä tiedoista pitäisi pystyä sinulle kertomaan? Missä? Miten? Missä muodosssa?

Tuleeko vielä jotain mieleen tai haluaisitko vielä lisätä jotain?

In all these examples, the personal information was handled by artificial intelligence. What do you think about this?

What be a difference if a human would handle those information? Why?

Which processes should be handled by Artificial Intelligence in such processes and which by a human? How much? Why?

Recently, you can experience more and more situations when you are interacting with artificial assistants, e.g. chatbots or automated phone calls. Imagine, those are using language natural enough to be indifferent with humans. What do you think about those?

Now imagine that you're calling somewhere, you have a good talk and you manage to do what intended, eg. book a visit. Only after the talk you get to know from a friend, that you were talking with AI. How would you feel?

What do you think about the use of AI to predict future health or mental problems, or social difficulties, like exclusion?

In which situations is this ok and when not ok? Why?

What would you like to or need to know about use cases like these?

What should your role be in the process (give consent, approve, review data?)

How should you be able to control? Why?

If we think about the processes, rules, algorithms and decisions made in those, and the data used, what are the things you would need to or like to know? Why?

How should all that information be presented to you? (When, where?)

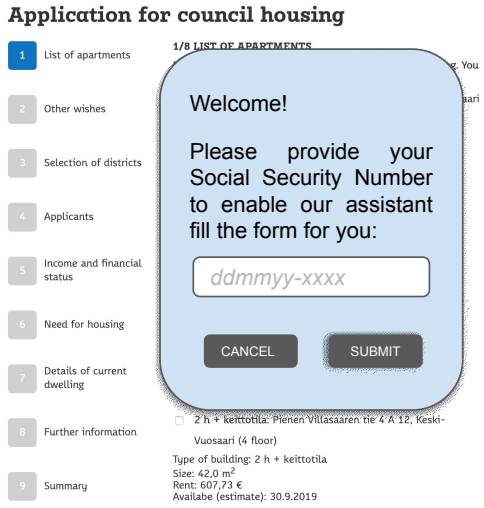
Appendix B

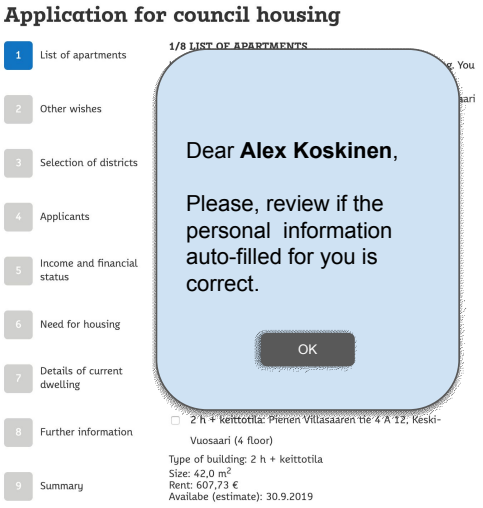
Interview cases

[redacted]

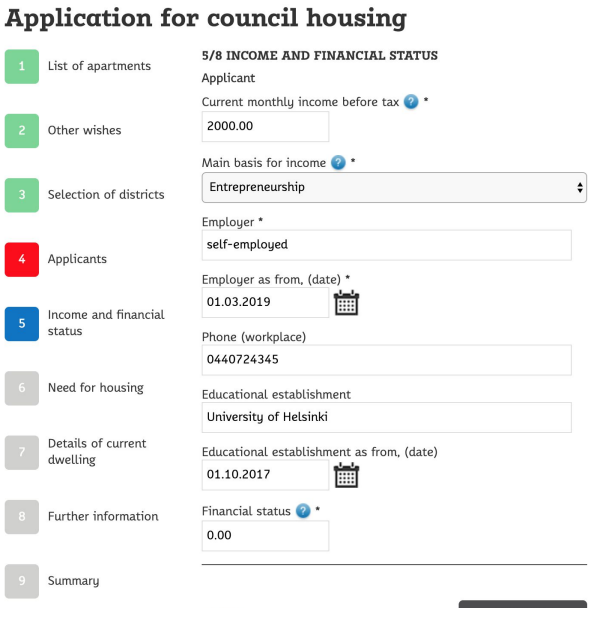
Case: Imagine that you're looking for a new flat. Having scarce funds for the rent, you're in need to apply for council housing in Helsinki. Below, you can see the application process.

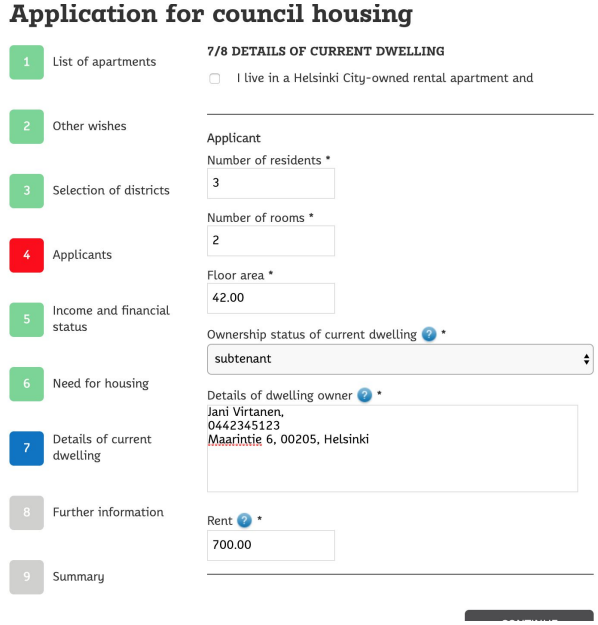
1. You fill your social security number and click submit.





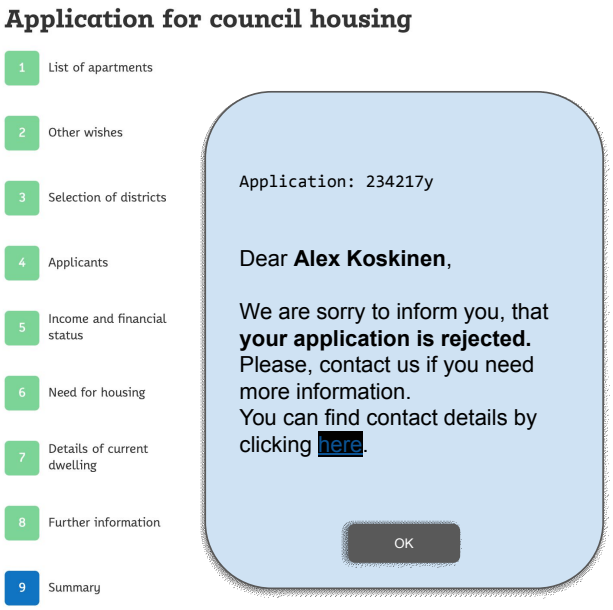
2. You go through the form which has auto-filled information about your current status of e.g. dwelling, income, family.





Case: Imagine that you're looking for a new flat. Having scarce funds for the rent, you're in need to apply for council housing in Helsinki. Below, you can see the application process.

3. After reviewing the form, you submit the application and the decision is presented immediately on the screen.



What are your feelings about this case?

Case: Let's imagine this situation. We are some years into the future. You have a grandmother who lives quite a distance from you. There have been some difficulties in the family and your grandmother has been living alone for quite some time and you haven't heard anything for a while.

Once, you receive a letter from "social care". It says:

Hello,

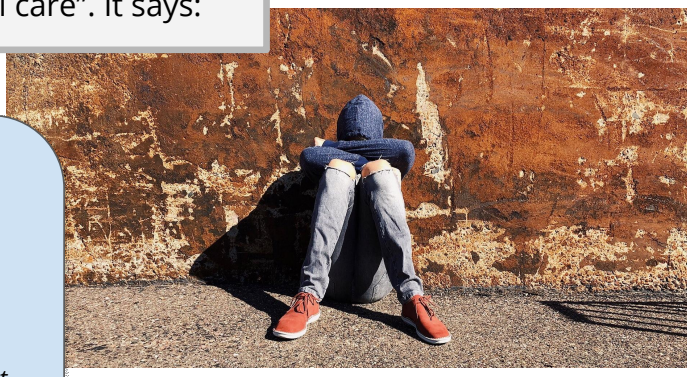
This letter concerns your grandmother.
Based on the information we have gathered, we have reason to suspect a danger of social exclusion in the case of your grandmother. We are not allowed to intervene at this point but would like to make you aware of the situation, as a next of kin.

Kind regards,
Social Care

What are your feelings about this case?

Case: Let's imagine this situation. We are some years into the future. You have a grandchild who lives quite a distance from you. There have been some difficulties in the family and your grandchild has been living alone for quite some time and you haven't heard anything for a while.

Once, you receive a letter from "social care". It says:



Hello,

This letter concerns your grandchild.
Based on the information we have gathered, we have reason to suspect a danger of social exclusion in the case of your grandchild. We are not allowed to intervene at this point but would like to make you aware of the situation, as a next of kin.

Kind regards,
Social Care

What are your feelings about this case?

[redacted]

se: Imagine that you have a
ld that just started primary
hool. One day, you got this
ssage from the school director:

Hello,

We are taking part in the new three-years project with a goal to assess the impact of our education on children. For this purpose, we would like to track your child information in this time. Please, let us know whether you are fine with your child contributing to the project. If so, please provide us the information whether your child use a mobile phone or a smartwatch, as we would use one of those for receiving information.

Thank you,
School Director.

What are your feelings about this case?

[redacted]

Case: Let's imagine that you are currently in the basic social assistance scheme and receiving benefits from Kela. (This means you have no income and couldn't cope without the benefits). You receive email from Kela which says:

Hello,

We have recently received information w
indicates that there might be a reason to c
your current benefits. We have reason to be
you have spent a considerable amount of m
recently. You have 5 days to claim adjustmer
the data before the benefits are cancelled.
You can review it here.

Thank you,
Kela

What are your feelings about this case?

Appendix C

Design workshop cases

[redacted]

[redacted]

[redacted]

qualitative answer (eg. school assignment to student).

Example: Imagine applying for a student or social housing. You fill in the form of current preferences, while more data is received from your housing history, education status, income. Then, you wait a bit and you are assigned a flat, as said, tailored to your needs.

[redacted]

[redacted]

[redacted]

[redacted]

Example: Finnish government wants to measure the impact of higher education in Finland on Finnish citizens' wellbeing and employability. Throughout a few years they gather information about fresh-graduates from different universities, e.g. their health profile, income, employment. After first year of the project, first results are published online.

[redacted]

[redacted]

[redacted]

[redacted]

Example: There is a specific desease repeating in your family. Moreover, the type of work you are doiing might increase the risk of getting it. The AI recognizes this risk and outputs the prediction: with the current information about you, you are likely to get this disease.

[redacted]

Appendix D

Personas

"Obviously, some new research in AI is going on and it's for the betterment of the future."

"I would be super happy to have a decision right away, because then I can take action faster."

a bot.

"If you're transparent about your service, then people will trust you in general. Even if they do not agree with your processes, then they can express their opinions on it, compared to if they didn't know your process.

[redacted]

[redacted]

[redacted]

[redacted]

"Monitoring our actions? Feels weird, like being a test rabbit."

"I see a lot of benefits in these, for example in terms of the child or social welfare."

[redacted]

[redacted]

Concern:

1. Surveillance.

"If I always know what's going on, then I feel ok. Otherwise, I would fear of what data is being used and for what purposes."

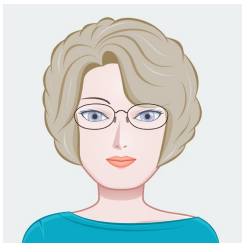
online are being tracked.

"You're either in or out of online world. Google knows everything."

"This would be pretty cruel if automation, not a person familiar with the situation, rejected the application.

not worry as long as is doing things legally.

"It's such a strange concept this artificial intelligence, it's probably doing things I can't comprehend"



[redacted]

[redacted]

[redacted]

"AI is nothing magic for me"

[redacted]

"The confidence comes from previous and general experience of government."

[redacted]

always interested to see the documentation of the service.

[redacted]

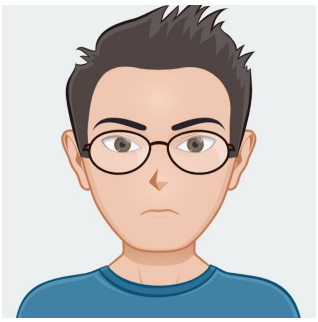
3. Data going in the wrong hands.
4. Bad impact on users.

[redacted]

documentation with curiosity.

3. Is interested in system properties, security and impact.

"Do we have the best people to keep the data protected in the public sector?"



[redacted]

[redacted]

[redacted]

"Using data and AI is a necessary evil"

[redacted]

policy towards AI.

"You should always be asked for permission to share your data, so it would not be automatically passed forward"

[redacted]

1. Misuse of data by the government.

[redacted]

the system.

"The expectation is that when you call a number and you talk to someone that sounds like you, it should be a human. If I discover that it was an automated process afterwards I would feel tricked."

Appendix E

Prototyped service

KAKSote
Alueellinen
terveydenhoito

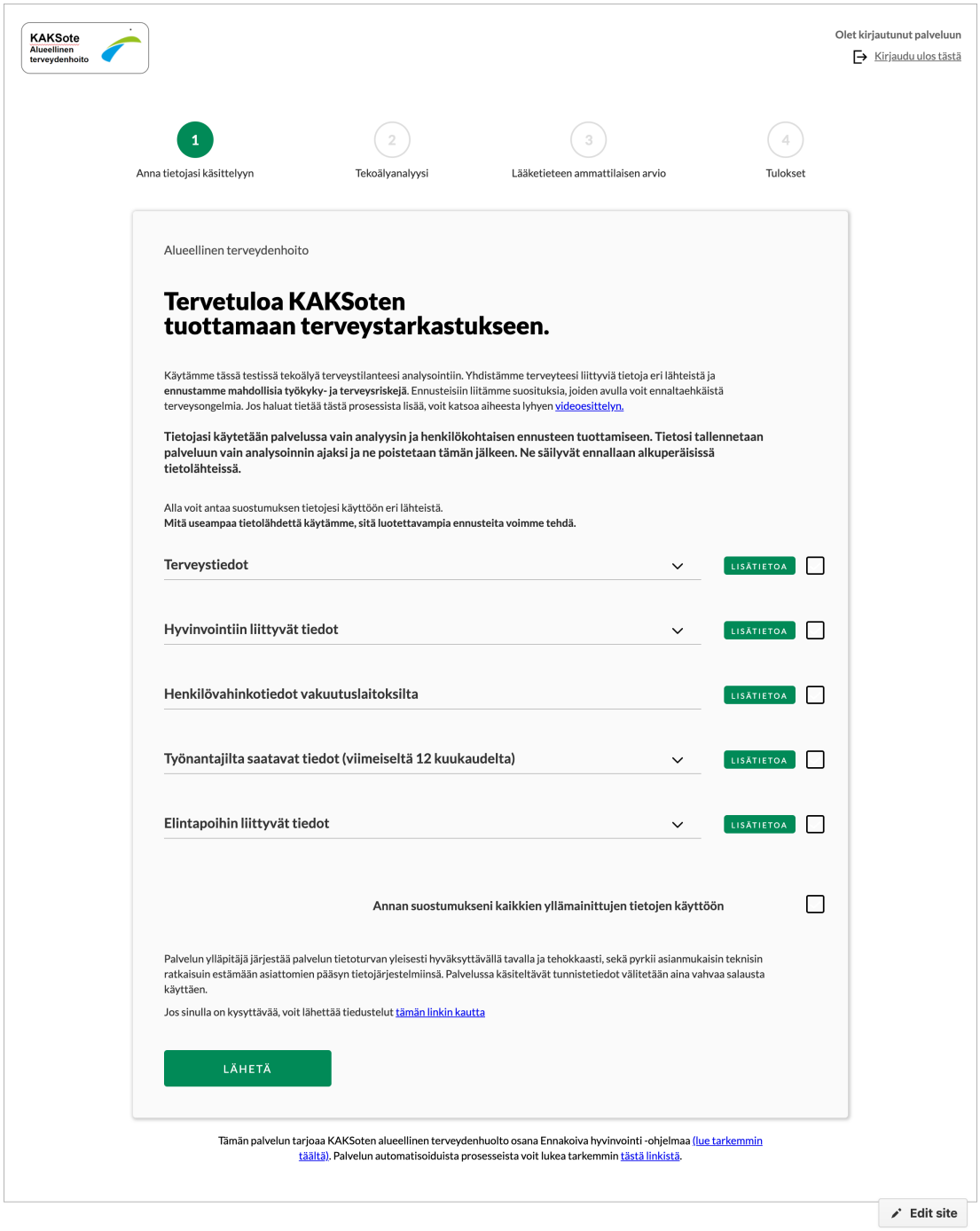
[redacted]

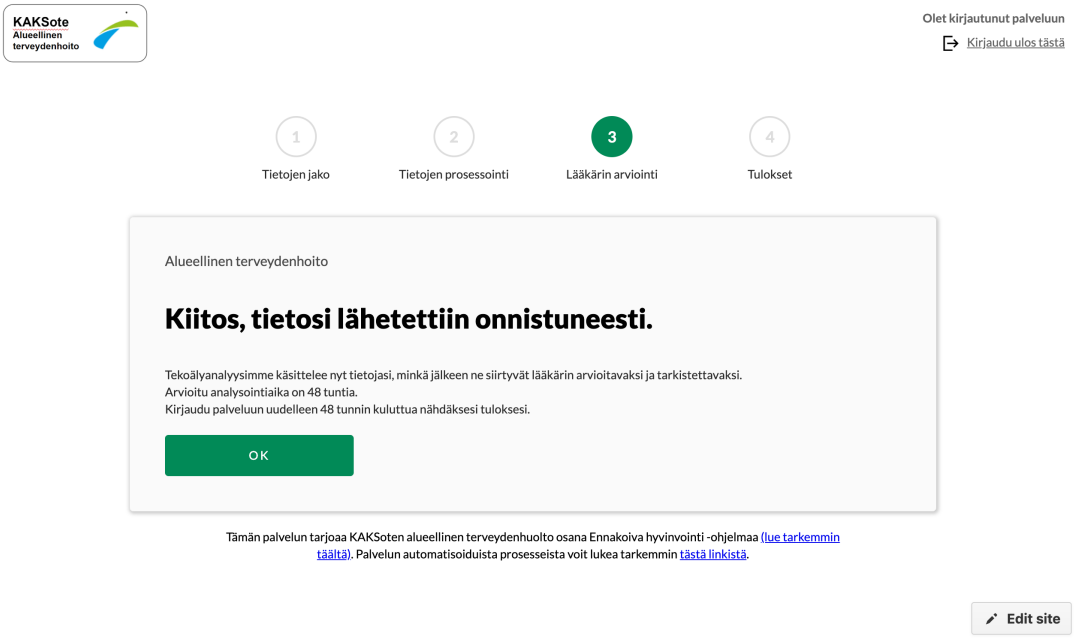
Tarkastukseen sisältyy tekoälyyn perustuvan terveysennusteen laatiminen tuleville työvuosillesi.

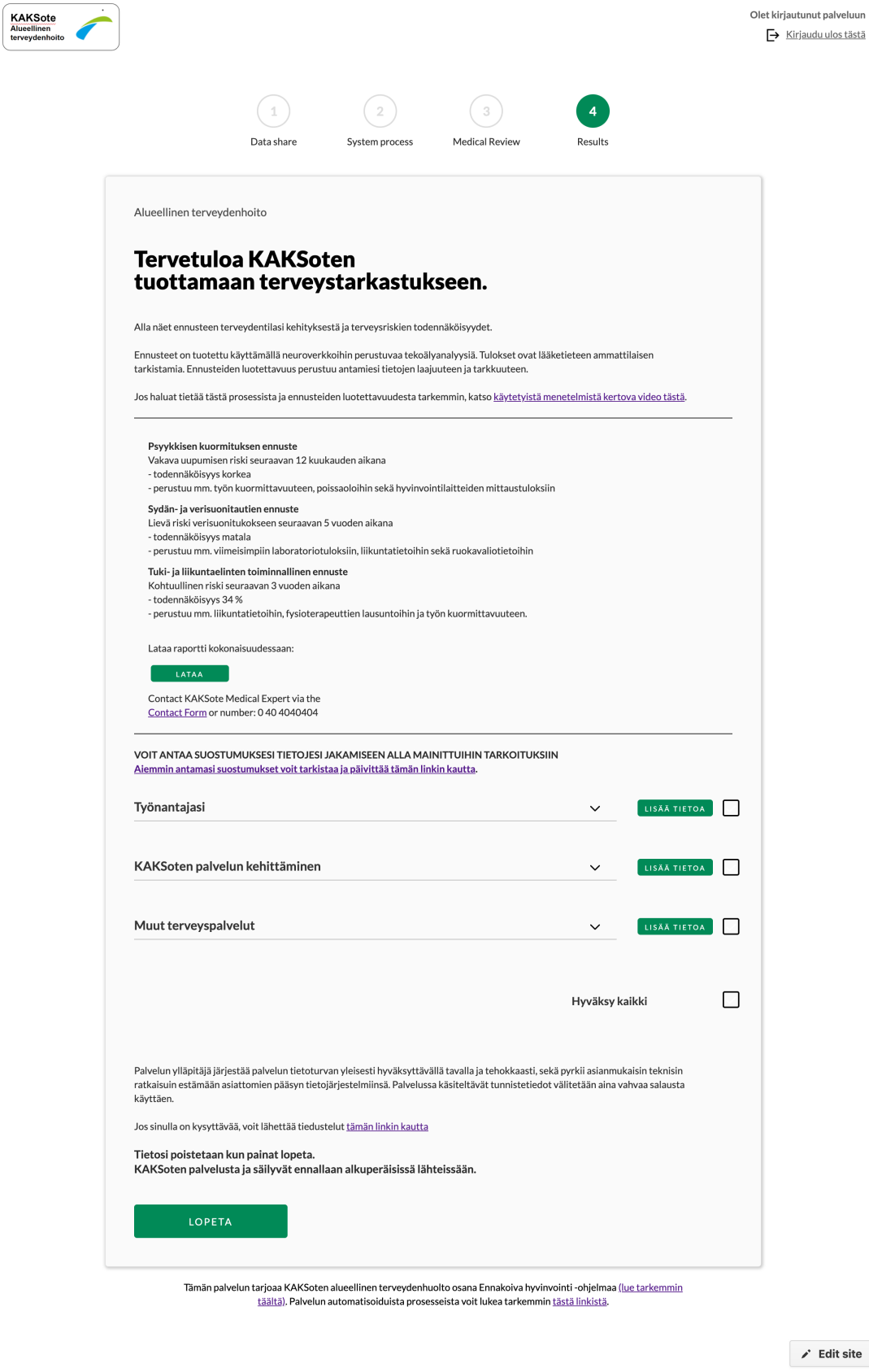
Voit perehtyä terveyennusteen taustoihin tarkemmin valitsemalla alta aiheen.

Teknologia	**Tiedot ja tietoturva**	**Suostumus ja tietosuoja**
Mitä tekoäly on?	Mitä tietoja ennuste käyttää?	Mistä tiedot saadaan?
Miten tekoälyä osaa ennustaa?	Miten tietosi on turvattu?	Kuka tietosi näkee?
Mitä algoritmeja käytämme?	Miten vältämme biasoitumiset?	Miten voit hallita tietojesi käyttöä?

KIRJAUDU PALVELUUN


KAKSote
Alueellinen terveydenhoito
Olet kirjautunut palveluun
Kirjaudu ulos tästä
1
Anna tietojasi käsittelyyn
2
Tekoälyanalyysi
3
Lääketieteen ammattilaisen arvio
4
Tulokset
Alueellinen terveydenhoito
Tervetuloa KAKSoten tuottamaan terveystarkastukseen.
Käytämme tässä testissä tekoälyä terveystilanteesi analysointiin. Yhdistämme terveyteesi liittyviä tietoja eri lähteistä ja ennustamme mahdollisia työkyky- ja terveysriskejä. Ennusteisiin liitämme suosituksia, joiden avulla voit ennaltaehkäistä terveysongelmia. Jos haluat tietää tästä prosessista lisää, voit katsoa aiheesta lyhyen videoesittelyn.
Tietojasi käytetään palvelussa vain analyysin ja henkilökohtaisen ennusteen tuottamiseen. Tietosi tallennetaan palveluun vain analysoinnin ajaksi ja ne poistetaan tämän jälkeen. Ne säilyvät ennallaan alkuperäisissä tietolähteissä.
Alla voit antaa suostumuksen tietojesi käyttöön eri lähteistä.
Mitä useampaa tietolähdettä käytämme, sitä luotettavampia ennusteita voimme tehdä.
Terveystiedot
LISÄTIETOA
Hyvinvointiin liittyvät tiedot
LISÄTIETOA
Henkilövahinkotiedot vakuutuslaitoksilta
LISÄTIETOA
Työnantajilta saatavat tiedot (viimeiseltä 12 kuukaudelta)
LISÄTIETOA
Elintapoihin liittyvät tiedot
LISÄTIETOA
Annan suostumukseni kaikkien yllämainittujen tietojen käyttöön
Palvelun ylläpitäjä järjestää palvelun tietoturvan yleisesti hyväksyttävällä tavalla ja tehokkaasti, sekä pyrkii asianmukaisin teknisin ratkaisuin estämään asiattomien pääsyn tietojärjestelmiinsä. Palvelussa käsiteltävät tunnistetiedot välitetään aina vahvaa salausta käyttäen.
Jos sinulla on kysyttävää, voit lähettää tiedustelut tämän linkin kautta
LÄHETÄ
Tämän palvelun tarjoaa KAKSoten alueellinen terveydenhuolto osana Ennakoiva hyvinvointi -ohjelmaa (lue tarkemmin täältä). Palvelun automatisoiduista prosesseista voit lukea tarkemmin tästä linkistä.
Edit site


KAKSote
Alueellinen terveydenhoito
Olet kirjautunut palveluun
Kirjaudu ulos tästä
1
Tietojen jako
2
Tietojen prosessointi
3
Lääkärin arviointi
4
Tulokset
Alueellinen terveydenhoito
Kiitos, tietosi lähetettiin onnistuneesti.
Tekoälyanalyysimme käsittelee nyt tietojasi, minkä jälkeen ne siirtyvät lääkärin arvioitavaksi ja tarkistettavaksi.
Arvioitu analysointiaika on 48 tuntia.
Kirjaudu palveluun uudelleen 48 tunnin kuluttua nähdäksesi tuloksesi.
OK
Tämän palvelun tarjoaa KAKSoten alueellinen terveydenhuolto osana Ennakoiva hyvinvointi -ohjelmaa (lue tarkemmin täältä). Palvelun automatisoiduista prosesseista voit lukea tarkemmin tästä linkistä.
Edit site


KAKSote
Alueellinen terveydenhoito
Olet kirjautunut palveluun
Kirjaudu ulos tästä
1
Data share
2
System process
3
Medical Review
4
Results
Alueellinen terveydenhoito
Tervetuloa KAKSoten tuottamaan terveystarkastukseen.
Alla näet ennusteen terveydentilasi kehityksestä ja terveysriskien todennäköisyydet.
Ennusteet on tuotettu käyttämällä neuroverkkoihin perustuvaa tekoälyanalyysiä. Tulokset ovat lääketieteen ammattilaisen tarkistamia. Ennusteiden luotettavuus perustuu antamiesi tietojen laajuuteen ja tarkkuuteen.
Jos haluat tietää tästä prosessista ja ennusteiden luotettavuudesta tarkemmin, katso käytetyistä menetelmistä kertova video tästä.
Psyykkisen kuormituksen ennuste
Vakava uupumisen riski seuraavan 12 kuukauden aikana
- todennäköisyys korkea
- perustuu mm. työn kuormittavuuteen, poissaoloihin sekä hyvinvointilaitteiden mittaustuloksiin
Sydän- ja verisuonitautien ennuste
Lievä riski verisuonitukokseen seuraavan 5 vuoden aikana
- todennäköisyys matala
- perustuu mm. viimeisimpiin laboratoriotuloksiin, liikuntatietoihin sekä ruokavaliotietoihin
Tuki- ja liikuntaelinten toiminnallinen ennuste
Kohtuullinen riski seuraavan 3 vuoden aikana
- todennäköisyys 34 %
- perustuu mm. liikuntatietoihin, fysioterapeuttien lausuntoihin ja työn kuormittavuuteen.
Lataa raportti kokonaisuudessaan:
LATAA
Contact KAKSote Medical Expert via the Contact Form or number: 0 40 4040404
VOIT ANTAA SUOSTUMUKSESI TIETOJESI JAKAMISEEN ALLA MAINITTUIHIN TARKOITUKSIIN
Aiemmin antamasi suostumukset voit tarkistaa ja päivittää tämän linkin kautta.
Työnantajasi
LISÄÄ TIETOA
KAKSoten palvelun kehittäminen
LISÄÄ TIETOA
Muut terveyspalvelut
LISÄÄ TIETOA
Hyväksy kaikki
Palvelun ylläpitäjä järjestää palvelun tietoturvan yleisesti hyväksyttävällä tavalla ja tehokkaasti, sekä pyrkii asianmukaisin teknisin ratkaisuin estämään asiattomien pääsyn tietojärjestelmiinsä. Palvelussa käsiteltävät tunnistetiedot välitetään aina vahvaa salausta käyttäen.
Jos sinulla on kysyttävää, voit lähettää tiedustelut tämän linkin kautta
Tietosi poistetaan kun painat lopeta.
KAKSoten palvelusta ja säilyvät ennallaan alkuperäisissä lähteissään.
LOPETA
Tämän palvelun tarjoaa KAKSoten alueellinen terveydenhuolto osana Ennakoiva hyvinvointi -ohjelmaa (lue tarkemmin täältä). Palvelun automatisoiduista prosesseista voit lukea tarkemmin tästä linkistä.
Edit site

Appendix F

Guidelines for trustworthy PS AI services

Guidelines for trustworthiness of public sector AI services

Below document contains guidelines for designers, developers and decision-makers of any public sector (PS) service that would use AI systems. Following such guidelines would build citizens' trust to the AI and public sector itself. Guidelines are the results of the empirical study done with Citizen Transparency project with Saidot in 2019[1] and extensive literature review, results of both are published in the *Designing guidelines for trustworthy AI public sector systems* Master Thesis.

Guidelines are divided into two sections. First section tells about how to achieve transparency of PS AI services, mainly what information, when and how to present for a citizen. Second one groups principles that should be included in design, development and operations of the service, as well as motivate for additional activities.

[1] https://www.espoo.fi/en-US/A_Consortium_of_Finnish_organisations_se(167195)

Information transparency

How, what and where should be shared in AI Public Sector services.

In the following section, we present guidelines for enabling transparency in Public Sector services that are using AI. Not only we present what information should be available for citizens, but also when and how to show those. We suggest that there are four different stages of citizen interactions with the AI PS service:

1. Informative stage
   Where a citizen can get informed about the service, its purpose and any details. That stage should be accessible at any point.
2. Application stage
   A citizen comes to the stage, when they decide to use the service. That's where the user can fill in any data, preferences or give consent.
3. Waiting stage
   In case, the service doesn't produce immediate results, there is a waiting stage where the user is unformed about the current state of the process.
4. Results stage
   Where the user is given results of the service.

At each stage, there is different information needed. Moreover, those information should be placed in the interface according to its priorities: the more important the information is, the earlier and more visible place it should have.

Information stage

- Process explained in better detail, therein exact role of human
- Technical documentation, therein performance of AI system, data quality, security of the system and data storage
- Whether the system is monitored, certified or audited
- (If possible) open sourced code
- Accountable person/organization
- Way of mitigating risks and bias
- Impact of the service on user and society, eg. visualisations of demographic
- Privacy statement

Application stage

Most important information - should be clearly visible:

- General description of the process, that AI will be used
- Basic information on the purpose
- What type of data will be collected about the person?
- What are the sources of these data?
- Security and privacy policy in short (what will happen with data, eg. whether they will be anonymized and shared with other organizations).

Important information, easily reachable:

- More detailed process, the role of AI and human in it
- More about the purpose and who benefits from the service
- Reason for using specific data
- Who will have access to the data?
- Whether data will be shared with other people/organisations?
- Basic information on the organisation involved in the process
- Basic information on the behaviour of the application
- A contact for inquiries

Where and how:

- Some suggest a person (eg. video clip), where the person would explain the process and the motives, in that case information should be also accessible by clear and easy text.
- The best way would be to do that via an online service or the application. From a citizen perspective, also contact via phone would be good.

Apart from the information, that stage should consist of a consent part. The citizen should be able to select whether they want to participate fully, only with limited, chosen data or not at all. If the data is planned to be shared with other organisations, the user also should consent for that. If possible, that could also involve an option of opting out from AI use or asking for human review.

Waiting stage

- Citizens should be able, either by an online service, or by a person of contact, change the permission (eg. participation) given at the consent. Eg. they should be

able to stop their participation, limit or increase the data they share. Opt out from ai, data sharing etc.

- If relevant (eg. while doing a longer research / impact assessment), citizens would like to get updates of the results.
- If possible, a person should be able to review the data being currently collected and possibly update some of those.

Results stage

By results we mean eg. finished prediction or decision.

Most important information, needed to be clearly visible:

- General description of the process
- Explanation
- Data that was used for the decision and its sources

Important information, easily reachable:

- Procedures or laws used in the service
- Performance of the used AI system (accuracy, likelihood)
- Data safety (what will happen with it now, if shared or deleted or reused)
- A redress contact
- When and where was the consent given
- If possible, link to review one's data

Increasing the experience:

- recommendations for what to do with given results
- guidance on how to read the results

Principles

What factors need to be provided by PS AI service for the citizen trust.

Below, there are principles presented that will help in building citizens' trust to AI and PS. They are divided into three sections, based on in which stage they should be addressed: service design, service development and operation and additional activities. Take into consideration that there exist also trust-building factors that are not controlled by project stakeholders, for example: already existing trust to the public sector, public opinion (word of mouth) and citizen personal experience.

Service design

Below guidelines suggest actions to be considered on the level of planning the service and its operations.

Give citizens control over their data

- Before any data is used, the owner of it should be asked for consent for using it.
- Citizens need to be given control of their personal data, that is they should be able to view, delete or change it.

Give citizens the option to interact with human

- In every service there should be an accessible way of contacting the human.
- The redress channel should be always available for citizens.
- Any sensitive or impactful results of the service, should rather be shared by a person.

Provide citizens with control over AI actions

- Citizens should have agency over AI systems, they should be able to revert or disable it.
- PS AI services should support citizens in making better decisions, rather than suverting it.

Make PS AI service efficient and beneficial

- PS AI service needs to bring in an efficient manner the benefit not only to the user, but also to the society and possibly the planet.

Respect citizens' rights

- PS AI service needs to obey citizens' freedom and constitutional procedures; and be compatible with cultural diversity, social norms and values.
- PS AI service cannot impose any lifestyle choices on citizens.

Choose accountable person

- There need to be people selected as accountable for any PS AI service operations.

Service development and operations

Below guidelines present factors needed to address during development and operation of the service.

Involve human in the process

- A human should be involved in the process of auditing, testing and monitoring of an AI system.
- A balance should be held between the autonomy of a human and AI. A human should be always in control of which decision and actions are done automatically and which manually.
- Human intervention should be enabled whenever needed.

Develop secure and reliable system

- Perform analysis of potential system failures, negative service impact and any other risks.
- Proactively prevent system failures.
- Ensure safety and reliability of the system, as well as security of data.

Provide privacy for citizens

- Ensuring the privacy of the citizen data.
- Do not use intimate or too old data.
- Storing only data that are required for the system, delete any other.
- Anonymize used data.

Address and mitigate possible bias

- Evaluate possible bias and discrimnation in the service.

- Mitigate biases, and if not possible, publicly share the information about those.
- Do not use demographic data, such as race or sex, for decisions.

Create accessible and good interface

- Make sure interfaces are following the accessibility standards and every citizen can use and understand them.
- Use simple, understandable language.
- Provide informative navigation through the service, make sure its behaviour is predictable.
- Make sure that the user will know at what stage of the process they and system are.
- Provide neutral, public-sector alike, aesthetic interface.

Follow other existing metrics

- Use frameworks or certification to facilitate audition or to ensure quality.

Ensure data quality

- Use only data from good and trusted sources.
- Do not use too old data.

Additional activities

Provide education

- Support educating citizens about AI via public events, like lectures, or accessible courses.
- Help in providing proper AI education in schools and universities, introducing both technical but also social aspects of AI.
- Educate people involved in development of PS AI services about the technology and connected ethics.

Keep citizen in the loop

- Support actions that help in providing citizens the feel that they are part of the society and their needs are heard.
- For example, organize public debates, dialogues or involving citizens in decision making.